



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ  
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ  
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ  
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΜΑΘΗΣΗΣ

# Detection of Hallucinations and Related Observable Overgeneration Mistakes

DIPLOMA THESIS

by

Natalia-Maria Grigoriadou

**Επιβλέπων:** Γεώργιος Στάμου  
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024





Εθνικό Μετσόβιο Πολυτεχνείο  
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών  
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών  
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

# Detection of Hallucinations and Related Observable Overgeneration Mistakes

## DIPLOMA THESIS

by

Natalia-Maria Grigoriadou

**Επιβλέπων:** Γεώργιος Στάμου  
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24<sup>η</sup> Οκτωβρίου, 2024.

.....  
Γεώργιος Στάμου  
Καθηγητής Ε.Μ.Π.

.....  
Αθανάσιος Βουλόδημος  
Επ. Καθηγητής Ε.Μ.Π.

.....  
Α.-Γ. Σταφυλοπάτης  
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024

.....  
**ΝΑΤΑΛΙΑ-ΜΑΡΙΑ ΓΡΗΓΟΡΙΑΔΟΥ**  
Διπλωματούχος Ηλεκτρολόγος Μηχανικός  
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © – All rights reserved Natalia-Maria Grigoriadou, 2024.  
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.





# Περίληψη

Τα τελευταία χρόνια, η εμφάνιση των μεγάλων γλωσσικών μοντέλων (LLMs) έχει φέρει επανάσταση στην Επεξεργασία Φυσικής Γλώσσας (NLP), αλλά έχει επίσης εγείρει κρίσιμες ανησυχίες σχετικά με την αξιοπιστία και την ακρίβεια των αποτελεσμάτων τους. Ένα από τα πιο πιεστικά ζητήματα είναι τα hallucination, όπου τα LLMs παράγουν πληροφορίες που είναι σημασιολογικά ασυνεπείς ή μη σχετικές με την είσοδο. Η αντιμετώπιση αυτής της πρόκλησης είναι υψίστης σημασίας για τη βελτίωση της αξιοπιστίας των LLMs, ιδίως σε εργασίες παραγωγής φυσικής γλώσσας (NLG).

Η παρούσα διατριβή διερευνά αποτελεσματικές μεθόδους για την ανίχνευση hallucination, εστιάζοντας σε ρυθμίσεις μαύρου κουτιού όπου οι εσωτερικές λειτουργίες του μοντέλου δεν είναι προσβάσιμες. Η έρευνα συμβάλλει στην ανίχνευση hallucination μέσω της συμμετοχής στο task 6 (SHROOM) του SemEval-2024, το οποίο περιλαμβάνει δυαδική ταξινόμηση σε Hallucination και not Hallucination σε task όπως η μηχανική μετάφραση, η παραγωγή ορισμών και η παραγωγή παραφράσεων. Αξιοποιώντας προ-εκπαιδευμένα μοντέλα που έχουν γίνει fine-tuned σε σύνολα δεδομένων ανίχνευσης hallucination και εξαγωγής συμπερασμάτων φυσικής γλώσσας (NLI), η εργασία αυτή επιτυγχάνει σημαντικές βελτιώσεις σε σχέση με τα βασικά συστήματα, με ακρίβεια που φτάνει περίπου το 80%.

Οι βασικές συνεισφορές αυτής της διατριβής περιλαμβάνουν την ανάπτυξη ενός ταξινομητή ψηφοφορίας συνόλου που συνδυάζει πολλαπλά μοντέλα για τη βελτίωση της ανίχνευσης hallucination, καθώς και μια σε βάθος ανάλυση των προκλήσεων που θέτουν οι διάφορες εργασίες NLG. Τα ευρήματα παρέχουν πολύτιμες γνώσεις σχετικά με τη φύση των hallucination στα LLMs και προσφέρουν ένα ισχυρό πλαίσιο για μελλοντικές προσπάθειες μετριασμού αυτού του εγγενούς προβλήματος.

**Λέξεις-κλειδιά** — Μεγάλα γλωσσικά μοντέλα (LLMs), ανίχνευση hallucination, Επεξεργασία Φυσικής Γλώσσας (NLP), Παραγωγή Φυσικής Γλώσσας (NLG), μοντέλα μαύρου κουτιού, ensemble voting classifier, fine-tuning, prompting, μηχανική μετάφραση, παραγωγή ορισμών, παραγωγή παραφράσεων, Natural Language Inference (NLI)



# Abstract

In recent years, the emergence of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP), but has also raised critical concerns regarding the trustworthiness and accuracy of their outputs. One of the most pressing issues is hallucination, where LLMs generate information that is semantically inconsistent or unrelated to the input. Addressing this challenge is paramount to improving the reliability of LLMs, especially in Natural Language Generation (NLG) tasks.

This thesis explores efficient methods for hallucination detection, focusing on black-box settings where the model’s internal workings are inaccessible. The research contributes to the detection of hallucinations through participation in the SemEval-2024 Task 6 (SHROOM), which involves binary classification of hallucinations across tasks such as machine translation, definition modeling, and paraphrase generation. By leveraging pre-trained models fine-tuned on hallucination detection and Natural Language Inference (NLI) datasets, this work achieves significant improvements over baseline systems, with accuracies reaching approximately 80%.

Key contributions of this thesis include the development of an ensemble voting classifier combining multiple models to enhance hallucination detection, as well as an in-depth analysis of the challenges posed by different NLG tasks. The findings provide valuable insights into the nature of hallucinations in LLMs and offer a robust framework for future efforts to mitigate this inherent issue.

**Keywords** — Large Language Models (LLMs), hallucination detection, Natural Language Processing (NLP), Natural Language Generation (NLG), black-box models, ensemble voting classifier, fine-tuning, prompting, machine translation, definition modeling, paraphrase generation, Natural Language Inference (NLI)



# Ευχαριστίες

Το έργο αυτό δεν θα ήταν δυνατό χωρίς την υποστήριξη και συμβολή πολλών ανθρώπων. Πρώτα απ' όλα, θα ήθελα να ευχαριστήσω τον επιβλέποντά μου, κ. Στάμου Γεώργιο για την καθοδήγηση του στην εκπόνηση αυτής της διπλωματικής. Επιπλέον θα ήθελα να ευχαριστήσω θερμά την Μαρία Λυμπεραίου και τον Γιώργο Φιλανδριανό για τη στενή συνεργασία, την υποστήριξη και την καθοριστική συμβολή τους από την αρχή μέχρι το τέλος της εργασίας αυτής. Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την στήριξή τους σε όλα τα χρόνια των σπουδών μου.

Ναταλία-Μαρία Γρηγοριάδου, Οκτώβριος 2024



# Contents

|                                               |           |
|-----------------------------------------------|-----------|
| <b>Contents</b>                               | <b>13</b> |
| <b>List of Figures</b>                        | <b>14</b> |
| <b>1 Εκτεταμένη Περίληψη στα Ελληνικά</b>     | <b>17</b> |
| 1.1 Εισαγωγή                                  | 17        |
| 1.2 Θεωρητικό Υπόβαθρο                        | 19        |
| 1.2.1 Μεγάλα γλωσσικά μοντέλα                 | 19        |
| 1.3 Προτεινόμενες Προσεγγίσεις                | 19        |
| 1.3.1 Σύνολο Δεδομένων                        | 19        |
| 1.3.2 Μετρικές                                | 20        |
| 1.3.3 Baseline System                         | 22        |
| 1.4 Μέθοδοι                                   | 22        |
| 1.5 Πειραματικό Μέρος                         | 25        |
| 1.5.1 Ομοιότητα προτάσεων                     | 25        |
| 1.5.2 COMET                                   | 37        |
| 1.5.3 Finetuning NLI μοντέλο                  | 37        |
| 1.5.4 Ταξινομητής Ψηφοφορίας                  | 40        |
| 1.5.5 Prompting                               | 40        |
| 1.6 Σύνοψη                                    | 45        |
| <b>2 Introduction</b>                         | <b>47</b> |
| <b>3 Related Work</b>                         | <b>49</b> |
| 3.1 Large Language Models                     | 49        |
| 3.2 Transformers                              | 50        |
| 3.3 LLMs Hallucinations                       | 51        |
| <b>4 Approach</b>                             | <b>55</b> |
| 4.1 Dataset                                   | 55        |
| 4.1.1 Data format                             | 55        |
| 4.1.2 Exploratory data analysis               | 56        |
| 4.2 Metrics                                   | 64        |
| 4.2.1 Accuracy                                | 64        |
| 4.2.2 Spearman Correlation                    | 64        |
| 4.3 Methods                                   | 65        |
| 4.3.1 Sentence Similarity                     | 65        |
| 4.3.2 COMET                                   | 65        |
| 4.3.3 Fine-tune NLI models                    | 66        |
| 4.3.4 Fine-tune hallucination detection model | 67        |
| 4.3.5 Voting Classifier                       | 68        |
| 4.3.6 Prompting                               | 68        |

|          |                                                   |            |
|----------|---------------------------------------------------|------------|
| <b>5</b> | <b>Experiments</b>                                | <b>71</b>  |
| 5.1      | Baseline System . . . . .                         | 72         |
| 5.2      | Sentence Similarity . . . . .                     | 72         |
| 5.2.1    | Calculate Distance . . . . .                      | 72         |
| 5.2.2    | Clustering . . . . .                              | 82         |
| 5.3      | COMET . . . . .                                   | 84         |
| 5.4      | NLI models . . . . .                              | 84         |
| 5.5      | Fine-tune hallucination detection model . . . . . | 87         |
| 5.6      | Voting Classifier . . . . .                       | 88         |
| 5.7      | Prompting . . . . .                               | 90         |
| 5.8      | Overall Results . . . . .                         | 94         |
| <b>6</b> | <b>Conclusion</b>                                 | <b>97</b>  |
| <b>7</b> | <b>Bibliography</b>                               | <b>99</b>  |
| <b>A</b> | <b>Thresholds Used for Sentence Similarity</b>    | <b>103</b> |

# List of Figures

|                                                                                                                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.3.1 Prompt template που χρησιμοποιήθηκε στο μοντέλο αναφοράς . . . . .                                                                                                                                       | 22 |
| 3.2.1 Transformer Architecture . . . . .                                                                                                                                                                       | 51 |
| 4.1.1 Distribution of per task samples in the initially released trial set. . . . .                                                                                                                            | 58 |
| 4.1.2 Distribution of unlabelled training samples per task in both model-agnostic and model-aware settings. . . . .                                                                                            | 59 |
| 4.1.3 Distribution of labeled validation samples per task in both model-agnostic and model-aware settings. . . . .                                                                                             | 59 |
| 4.1.4 Distribution of validation labels in both model-agnostic and model-aware settings. . . . .                                                                                                               | 60 |
| 4.1.5 Distribution of hallucination probability (majority voting among human annotators' labeling) in both model-agnostic and model-aware settings in the validation set. . . . .                              | 60 |
| 4.1.6 Distribution of hallucination probability in each validation label ('Hallucination' vs 'Not Hallucination'). Annotators significantly agree on whether a sample contains a hallucination or not. . . . . | 61 |
| 4.1.7 Distribution of labeled test samples per task in both model-agnostic and model-aware settings. . . . .                                                                                                   | 61 |
| 4.1.8 Distribution of test labels in both model-agnostic and model-aware settings. . . . .                                                                                                                     | 62 |
| 4.1.9 Distribution of hallucination probability (majority voting among human annotators' labeling) in both model-agnostic and model-aware settings in the test set. . . . .                                    | 62 |
| 4.1.10 Distribution of hallucination probability in each test label ('Hallucination' vs 'Not Hallucination'). Annotators significantly agree on whether a sample contains a hallucination or not. . . . .      | 63 |
| 5.1.1 Prompt template used in the baseline system . . . . .                                                                                                                                                    | 72 |
| 5.8.1 Percentage of errors per label in the model-agnostic and aware test set. . . . .                                                                                                                         | 94 |
| 5.8.2 Percentage of errors per task in the model-agnostic and aware test set. . . . .                                                                                                                          | 95 |
| 5.8.3 Percentage of errors per p('Hallucination') in the model-agnostic and aware test set. . . . .                                                                                                            | 95 |



# Chapter 1

## Εκτεταμένη Περίληψη στα Ελληνικά

### Contents

|            |                                   |           |
|------------|-----------------------------------|-----------|
| <b>1.1</b> | <b>Εισαγωγή</b>                   | <b>17</b> |
| <b>1.2</b> | <b>Θεωρητικό Υπόβαθρο</b>         | <b>19</b> |
| 1.2.1      | Μεγάλα γλωσσικά μοντέλα           | 19        |
| <b>1.3</b> | <b>Προτεινόμενες Προσεγγίσεις</b> | <b>19</b> |
| 1.3.1      | Σύνολο Δεδομένων                  | 19        |
| 1.3.2      | Μετρικές                          | 20        |
| 1.3.3      | Baseline System                   | 22        |
| <b>1.4</b> | <b>Μέθοδοι</b>                    | <b>22</b> |
| <b>1.5</b> | <b>Πειραματικό Μέρος</b>          | <b>25</b> |
| 1.5.1      | Ομοιότητα προτάσεων               | 25        |
| 1.5.2      | COMET                             | 37        |
| 1.5.3      | Finetuning NLI μοντέλο            | 37        |
| 1.5.4      | Ταξινομητής Ψηφοφορίας            | 40        |
| 1.5.5      | Prompting                         | 40        |
| <b>1.6</b> | <b>Σύνοψη</b>                     | <b>45</b> |

### 1.1 Εισαγωγή

Τα τελευταία χρόνια, ο τομέας της Επεξεργασίας Φυσικής Γλώσσας (NLP) έχει μεταμορφωθεί σημαντικά από την εμφάνιση και την ευρεία υιοθέτηση των μεγάλων γλωσσικών μοντέλων (LLM). Τα μοντέλα αυτά όχι μόνο έφεραν επανάσταση στον τρόπο με τον οποίο προσεγγίζονται οι εργασίες φυσικής γλώσσας, αλλά δημιούργησαν και νέες ανησυχίες σχετικά με την αξιοπιστία τους και την ακρίβεια των αποτελεσμάτων που παράγουν. Ένα από τα πιο πιεστικά ζητήματα που αντιμετωπίζει η κοινότητα NLP είναι το φαινόμενο των hallucinations στα LLMs, κατά το οποίο τα μοντέλα παράγουν πληροφορίες που δεν βασίζονται στα δεδομένα εισόδου ή στην πραγματικότητα. Το ζήτημα αυτό έχει συγκεντρώσει σημαντική προσοχή στην ερευνητική κοινότητα, καθώς επηρεάζει άμεσα την αξιοπιστία και τη χρησιμότητα των συστημάτων που βασίζονται σε LLMs σε ένα ευρύ φάσμα εφαρμογών. Παρά τις εντυπωσιακές δυνατότητες των LLMs, τα hallucinations αποτελούν σημαντική πρόκληση, καθώς υπονομεύουν την αξιοπιστία των συστημάτων που έχουν σχεδιαστεί για να βοηθούν στη λήψη αποφάσεων, στην παραγωγή περιεχομένου και άλλα.

Πρόσφατες μελέτες έχουν υποδείξει ότι τα hallucinations δεν είναι απλώς τυχαία ή αποφεύξιμα σφάλματα, αλλά αντίθετα είναι εγγενή στη φύση των LLMs, δεδομένου του πιθανολογικού τους υποβάθρου. Τα ευρήματα αυτά δείχνουν ότι τα hallucinations μπορεί να είναι αναπόφευκτα σε κάποιο βαθμό, ανεξάρτητα από την πολυπλοκότητα της αρχιτεκτονικής του μοντέλου. Ως εκ τούτου, η αντιμετώπιση του ζητήματος των hallucinations, ιδίως στο πλαίσιο των εργασιών παραγωγής φυσικής γλώσσας (NLG), έχει γίνει ένας κρίσιμος τομέας εστίασης. Ωστόσο,

παρά τον αυξανόμενο όγκο ερευνών επί του θέματος, εξακολουθεί να υπάρχει αξιοσημείωτη έλλειψη συνόλων δεδομένων και κριτηρίων αναφοράς ειδικά σχεδιασμένων για την αξιολόγηση και τον μετριασμό των hallucination σε εργασίες NLG.

Σε απάντηση αυτής της ανάγκης, το task 6 του SemEval-2024 (SHROOM: Shared-task on Hallucinations and Related Observable Overgeneration Mistakes) δημιουργήθηκε για να αντιμετωπίσει την πρόκληση του εντοπισμού σημασιολογικά μη σχετικών παραγωγών γλωσσικών κειμένων σε εργασίες NLP. Αυτή η εργασία περιλαμβάνει διάφορους τομείς με προκλήσεις, όπως η μετάφραση, ο ορισμός λέξεων ή φράσεων και η παραγωγή παραφράσεων. Αυτό που καθιστά το SHROOM ιδιαίτερα απαιτητικό είναι η συμπερίληψη τόσο γνωστών όσο και άγνωστων αρχιτεκτονικών μοντέλου, που σπρώχνει τα όρια των τρεχουσών στρατηγικών ανίχνευσης hallucination και απαιτεί την ανάπτυξη προσεγγίσεων που είναι ανθεκτικές σε διαφορετικά επίπεδα διαφάνειας του μοντέλου.

Η παρούσα διπλωματική αποσκοπεί στη διερεύνηση αποτελεσματικών και ευρέως εφαρμόσιμων στρατηγικών για την ανίχνευση hallucination, ιδίως σε ρυθμίσεις black-box, όπου οι εσωτερικές λειτουργίες του μοντέλου που παράγει το κείμενο δεν είναι προσβάσιμες. Ακόμα και στις περιπτώσεις όπου το υποκείμενο μοντέλο είναι γνωστό, όπως στον διαγωνισμό SHROOM, τηρούμε μια αυστηρή black-box προσέγγιση, που σημαίνει ότι δεν αναγεννάμε τις εξόδους από τα μοντέλα, αλλά αντίθετα επικεντρωνόμαστε αποκλειστικά στην ανίχνευση hallucination στις υπάρχουσες εξόδους.

Για το σκοπό αυτό, αξιοποιούμε προ-εκπαιδευμένα μοντέλα που ενσωματώνουν γνώση των σημασιολογικών σχέσεων που σχετίζονται με τα hallucinations. Μέσω της τελειοποίησης αυτών των μοντέλων σε επισημειωμένα παραδείγματα από το σύνολο δεδομένων SHROOM, επιτυγχάνουμε κατά προσέγγιση ακρίβεια 80% στην ανίχνευση hallucination, η οποία ξεπερνά σημαντικά την αρχική ακρίβεια 74,5% του baseline μοντέλου που δόθηκε από τους διοργανωτές του task, η οποία επιτεύχθηκε με τη χρήση του Mistral μοντέλου ανοικτού κώδικα.

Η παρούσα εργασία συνεισφέρει σε πολλά βασικά σημεία:

1. Πραγματοποιούμε fine-tuning σε μοντέλα προ-εκπαιδευμένα σε σύνολα δεδομένων ανίχνευσης hallucinations και εξαγωγής συμπερασμάτων φυσικής γλώσσας (NLI), τα οποία είναι σημασιολογικά ευθυγραμμισμένα με τις προκλήσεις που παρουσιάζει το SHROOM.
2. Τα εκλεπτυσμένα μοντέλα ενσωματώνονται σε ένα πλαίσιο ταξινομητή ψηφοφορίας (Voting Classifier), το οποίο αποδίδει ανταγωνιστική ακρίβεια ανίχνευσης.
3. Σχεδιάζουμε μια μέθοδο prompting που ενισχύει την ερμηνευσιμότητα της ανίχνευσης hallucinations καθοδηγώντας τα μοντέλα να εστιάζουν σε βασικές σημασιολογικές ασυνέπειες στις παραγόμενες εξόδους.
4. Οι μετρικές ομοιότητας προτάσεων, που προέρχονται από ενσωματώσεις που παράγονται από προ-εκπαιδευμένα μοντέλα, αξιοποιούνται για την ποσοτικοποίηση και την ανίχνευση hallucinations μέσω της σύγκρισης των παραγόμενων εξόδων με τα δεδομένα εισόδου.
5. Όλος ο πειραματισμός που διεξάγεται σε αυτή την εργασία έχει σχεδιαστεί ώστε να είναι αποδοτικός από άποψη χρόνου και υπολογιστικής απόδοσης, τηρώντας τους περιορισμούς ενός πλήρως black-box περιβάλλοντος.
6. Παρέχεται λεπτομερής ανάλυση των αποτελεσμάτων, αναλυτικά ανά εργασία, προσφέροντας πληροφορίες τόσο για την επιτυχή ανίχνευση hallucination όσο και για τις προκλήσεις που θέτουν οι αποτυχημένες περιπτώσεις, συμβάλλοντας στη βαθύτερη κατανόηση της φύσης των hallucination στις εξόδους LLM.

Συνοψίζοντας, η παρούσα έρευνα όχι μόνο προάγει την κατάσταση της τεχνολογίας στην ανίχνευση hallucinations για LLM, αλλά παρέχει επίσης τα θεμέλια για μελλοντικές εργασίες που αποσκοπούν στον μετριασμό αυτού του εγγενούς ζητήματος σε εργασίες παραγωγής φυσικής γλώσσας. Με τη διερεύνηση της διασταύρωσης της αρχιτεκτονικής του μοντέλου, της εξειδίκευσης της εργασίας και των μετρικών αξιολόγησης, συμβάλλουμε στην ανάπτυξη πιο αξιόπιστων και αξιόπιστων συστημάτων που βασίζονται σε LLM.

## 1.2 Θεωρητικό Υπόβαθρο

### 1.2.1 Μεγάλα γλωσσικά μοντέλα

Τα γλωσσικά μοντέλα (*LMs*) είναι υπολογιστικά πλαίσια σχεδιασμένα για την κατανόηση και την παραγωγή της ανθρώπινης γλώσσας. Πιο συγκεκριμένα, ένα γλωσσικό μοντέλο ορίζει μια κατανομή πιθανοτήτων πάνω σε ακολουθίες λέξεων, επιτρέποντας την πρόβλεψη της πιθανότητας αυτών των ακολουθιών ή τη δημιουργία νέου κειμένου με βάση μια είσοδο.

Ιστορικά, η θεμελιώδης προσέγγιση για την πιθανολογική μοντελοποίηση της γλώσσας ήταν τα *n-gram models*. Αυτά τα μοντέλα ακολουθούν την υπόθεση της αλυσίδας Markov, όπου η πιθανότητα της επόμενης λέξης σε μια ακολουθία εξαρτάται από ένα σταθερό παράθυρο προηγούμενων λέξεων. Για παράδειγμα, ένα μοντέλο bigram θεωρεί μία προηγούμενη λέξη, ένα trigram δύο, και ένα μοντέλο n-gram θεωρεί  $n - 1$  προηγούμενες λέξεις. Η πιθανότητα μιας ακολουθίας  $w_1, w_2, \dots, w_n$  σε ένα μοντέλο bigram αναπαρίσταται ως εξής:

$$P(w_1, w_2, \dots, w_n) = P(w_2|w_1) \cdot P(w_3|w_2) \cdots P(w_n|w_{n-1}) \quad (1.2.1)$$

Εδώ, η υπό συνθήκη πιθανότητα  $P(w_k|w_{k-1})$  εκτιμάται από τη σχετική συχνότητα των εμφανίσεων όπου  $w_{k-1}$  ακολουθείται από  $w_k$  στο σώμα εκπαίδευσης. Η εκπαίδευση ενός μοντέλου n-gram περιλαμβάνει την εκτίμηση αυτών των πιθανοτήτων από σώματα κειμένων, συνήθως σε μία ή περισσότερες γλώσσες.

Ένας βασικός περιορισμός των μοντέλων n-gram είναι η αδυναμία τους να αποδώσουν μη μηδενικές πιθανότητες σε ακολουθίες που δεν συναντώνται κατά την εκπαίδευση, δεδομένου του άπειρου εύρους έγκυρων προτάσεων που μπορεί να εκφράσει μια γλώσσα. Για την αντιμετώπιση αυτού του ζητήματος έχουν προταθεί διάφορες τεχνικές εξομάλυνσης.

Η έλευση των νευρωνικών δικτύων εισήγαγε την εποχή των νευρωνικών γλωσσικών μοντέλων, ξεκινώντας με απλά νευρωνικά γλωσσικά μοντέλα [4]. Ένα νευρωνικό γλωσσικό μοντέλο feed-forward είναι ένα απλό νευρωνικό δίκτυο που δέχεται έναν σταθερό αριθμό προηγούμενων λέξεων ως είσοδο και εξάγει μια κατανομή πιθανότητας για τις πιθανές επόμενες λέξεις. Παρόμοια με ένα μοντέλο n-gram, ένα νευρωνικό γλωσσικό μοντέλο feed-forward προβλέπει την πιθανότητα μιας λέξης σε συνάρτηση με τις  $n - 1$  προηγούμενες λέξεις.

Μια σημαντική διάκριση μεταξύ των μοντέλων n-gram και των νευρωνικών γλωσσικών μοντέλων feed-forward έγκειται στην αναπαράσταση των ακολουθιών λέξεων. Ενώ τα μοντέλα n-gram χρησιμοποιούν διακριτές ταυτότητες λέξεων, τα νευρωνικά γλωσσικά μοντέλα feed-forward αξιοποιούν την ενσωμάτωση λέξεων, συνεχείς διανυσματικές αναπαραστάσεις των λέξεων. Αυτό έχει ως αποτέλεσμα ανώτερες επιδόσεις γενίκευσης, ιδίως για αθέατες ακολουθίες λέξεων κατά τη διάρκεια της δοκιμής.

Αν και τα νευρωνικά γλωσσικά μοντέλα feed-forward σηματοδότησαν σημαντική πρόοδο στη μοντελοποίηση γλωσσών, τα σύγχρονα νευρωνικά γλωσσικά μοντέλα βασίζονται σε πιο εξελιγμένες αρχιτεκτονικές, όπως τα αναδρομικά δίκτυα και οι μετασχηματιστές.

*Recurrent neural network (RNN)* Τα γλωσσικά μοντέλα [36] επεξεργάζονται ακολουθίες εισόδου διαδοχικά, αναλύοντας μία λέξη κάθε φορά. Στόχος τους είναι να προβλέψουν την επόμενη λέξη με βάση την τρέχουσα λέξη και την προηγούμενη κρυφή κατάσταση. Σε αντίθεση με τα μοντέλα n-gram, τα RNN μπορούν να συλλάβουν εξαρτήσεις μεγάλης εμβέλειας στην ακολουθία εισόδου, καθώς δεν βασίζονται σε ένα σταθερού μεγέθους πλαίσιο. Η κρυφή κατάσταση στα RNNs μπορεί, θεωρητικά, να αντιπροσωπεύει πληροφορίες από όλες τις προηγούμενες λέξεις της ακολουθίας, αντιμετωπίζοντας το περιορισμένο πλαίσιο των μοντέλων n-gram και των νευρωνικών δικτύων feed-forward.

## 1.3 Προτεινόμενες Προσεγγίσεις

### 1.3.1 Σύνολο Δεδομένων

Με γνώμονα τις επερχόμενες προκλήσεις στο πεδίο του NLG, το σύνολο δεδομένων SHROOM επικεντρώνεται στα κυρίαρχα ζητήματα των μοντέλων που παράγουν γραμματικά ορθά αλλά ανακριβή (λανθασμένα ή μη υποστηριζόμενα) αποτελέσματα. Στους συμμετέχοντες ανατέθηκε η δυαδική ταξινόμηση για τον εντοπισμό

hallucination σε απαντήσεις μοντέλων, είτε όταν είναι γνωστό το μοντέλο είτε όχι (*model-aware* και *model-agnostic*). Η εργασία περιλαμβάνει τρεις τομείς NLG - μοντελοποίηση ορισμού (DM), μετάφραση (MT) και παραγωγή παραφράσεων (PG) - παρέχοντας εισόδους, αναφορές και εξόδους για δυαδική ταξινόμηση. Το σύνολο ανάπτυξης περιλαμβάνει ετικέτα ταξινόμησης από πολλαπλούς ανθρώπους, καθιερώνοντας μια "χρυσή" ετικέτα πλειοψηφείας.

**Μορφή Δεδομένων** Σε όλες τις περιπτώσεις τα δεδομένα ακολουθούν συγκεκριμένη μορφή:

- *src*: η είσοδος που δόθηκε στο μοντέλο
- *hyp*: η έξοδος που δόθηκε από το μοντέλο
- *tgt*: η σωστή απάντηση που περιμέναμε να δώσει το μοντέλο
- *ref*: ποιο από το 'tgt' ή 'src' (ή και τα δύο) περιέχει την απαραίτητη πληροφορία για την ταξινόμηση
- *task*: ποιος από τους τρεις τομείς του NLG (MT, DM, PG) ήταν ο σκοπός της χρήσης του μοντέλου
- *model*: το μοντέλο που χρησιμοποιήθηκε (στην περίπτωση του model-agnostic το συγκεκριμένο είναι άδειο).

Αρχικά, δόθηκαν 80 δείγματα με τις ετικέτες τους (trial set), ακολουθούμενα από δεδομένα εκπαίδευσης αποτελούμενα από 30k model-agnostic και 30k model-aware δείγματα χωρίς τις ετικέτες (train set). Τέλος, δόθηκε το σύνολο επικύρωσης (validation set) με 499 και 501 δείγματα, με τις ετικέτες τους, από model-agnostic και model-aware αντίστοιχα, και το τελικό σύνολο αξιολόγησης (test set) περιλαμβάνει 1500 δείγματα model-aware με ετικέτες και 1500 δείγματα model agnostic με ετικέτες.

Στα σύνολα που περιλαμβάνονται οι ετικέτες, δίνονται και επιπλέον πληροφορίες. Συγκεκριμένα δίνονται:

- *labels*: μια λίστα από 5 ετικέτες ('Hallucination' ή 'Not Hallucination') από 5 διαφορετικούς ανθρώπους
- *label*: η τελική ετικέτα που προκύπτει από την πλειοψηφία της λίστας των labels
- *p(Hallucination:)* το οποίο δηλώνει την πιθανότητα hallucination που προκύπτει ως ποσοστό των ετικετών από την λίστα labels που χαρακτηρίσαν ως hallucination το δείγμα

Αξίζει να σημειωθεί ότι στην πρώτη φάση του διαγωνισμού, μας δόθηκαν μόνο τα train, validation και trial σύνολα δεδομένων ενώ στην δεύτερη φάση μας δόθηκε και το test set χωρίς τις ετικέτες. Επομένως τα πειράματα διεξήχθησαν στο validation set και αργότερα, μετά την λήξη του διαγωνισμού που δόθηκαν και οι υπόλοιπες ετικέτες, στο test set.

Στον πίνακα 1.1 δείχνουμε κάποια δείγματα από το training σύνολο δεδομένων. Σχετικά με το task της μετάφρασης (MT) ανιχνεύσαμε ένα πλήθος γλωσσών μεταξύ των οποίων τα Ρώσικα, Αραβικά, Κινέζικα, Yorùbá, Telugu, Tsonga, Uzbek, Sinhalese, Quechuan, Mizo και άλλων. Πληροφορίες σχετικά με την γλώσσα δεν δόθηκαν, οπότε οι πληροφορίες αυτές διερευνήθηκαν χειροκίνητα μέσω του *src*.

### 1.3.2 Μετρικές

Προκειμένου να αξιολογήσουμε τα αποτελέσματα των πειραμάτων χρησιμοποιήθηκαν δύο μετρικές, σύμφωνα πάντα με τις οδηγίες του Semeval διαγωνισμού:

- **Accuracy** for the binary labels ('Hallucination' and 'Not Hallucination')
- **Spearman Correlation** for the percentage of Hallucination ('p(Hallucination)').

**Accuracy** Accuracy [1] είναι μια μετρική που χρησιμοποιείται για να δείξει πόσο συχνά το μοντέλο μηχανικής μάθησης κάνει σωστές προβλέψεις. Υπολογίζεται διαιρώντας τον αριθμό των σωστών προβλέψεων με τον αριθμό των συνολικών προβλέψεων που έγιναν, όπως φαίνεται και στην επόμενη εξίσωση:

$$Accuracy = \frac{\# \text{ΣωστεςΠροβλεψεις}}{\# \text{ΣυνολικεςΠροβλεψεις}} \quad (1.3.1)$$

| Model-agnostic        |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Machine Translation   | 'hyp': "Don't worry, it's only temporary.", 'tgt': "Don't worry. It's only temporary.", 'src': 'He волнуется. Это только временно.', 'ref': 'either', 'task': 'MT', 'model': "                                                                                                                                                                                                                                                                          |
| Definition modelling  | 'hyp': '(uncountable) The quality of being oronymy; the state of being oronymy.', 'tgt': 'The nomenclature of mountains, hills and other geographic rises.', 'src': 'An ancient survival in Turkish <define> oronymy </define> is quite possible , but I have not found Nihan Dag on the relevant sheets of the 1 : 200,000 map of Turkey , which are very detailed in matters of oronymy ;', 'ref': 'tgt', 'task': 'DM', 'model': "                    |
| Definition modelling  | 'hyp': '(intransitive, obsolete) To make a magazin of; to compose a magazin.', 'tgt': '(colloquial) The act of editing or writing for a magazine.', 'src': "Thus , though Byron is gone after his Don Juan — Scott and Southey out of the rhyme department — Wordsworth stamp - mastering — Coleridge 's poetry in abeyance — Crabbe mute as a fish - Campbell and Wilsont merely <define> magazing </define> ", 'ref': 'tgt', 'task': 'DM', 'model': " |
| Paraphrase Generation | 'hyp': 'You got something for me, huh?', 'tgt': ", 'src': 'Got something for me?', 'ref': 'src', 'task': 'PG', 'model': "                                                                                                                                                                                                                                                                                                                               |
| Model-aware           |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Machine Translation   | 'hyp': 'Our Mailoamiris of the System of Treatment of Ulilaes have created a place for these little ones.', 'tgt': 'We perceive the Foster Care System to be a safety zone for these children.', 'src': 'Maamiris tayo a ti Sistema iti Panangtaripato kadagiti Ulila ket natalged a lugar para kadagitoy nga ubbing.', 'ref': 'either', 'task': 'MT', 'model': 'facebook/nllb-200-distilled-600M'                                                      |
| Definition modeling   | 'hyp': 'To be obsequiously interested in .', 'tgt': '( usually followed by over or after ) To fuss over something adoringly ; to be infatuated with someone .', 'src': "Sarah mooned over sam 's photograph for months . What is the meaning of moon ? ", 'ref': 'tgt', 'task': 'DM', 'model': 'ltg/flan-t5-definition-en-base'                                                                                                                         |
| Paraphrase Generation | 'hyp': "Mr Barros Moura's report looks to the future in my opinion.", 'tgt': ", 'src': 'In my opinion, the most important element of the report by Mr Barros Moura is that it looks to the future.', 'ref': 'src', 'task': 'PG', 'model': 'tuner007/pegasus_paraphrase'                                                                                                                                                                                 |

Table 1.1: Δείγματα από το training σύνολο δεδομένων

**Spearman Correlation** Η μετρική συσχέτισης Spearman [9] είναι ένα στατιστικό μέτρο που αξιολογεί την ισχύ και την κατεύθυνση μιας μονότονης σχέσης μεταξύ δύο μεταβλητών συγκρίνοντας τις τιμές κατάταξής τους. Η συσχέτιση Spearman είναι χρήσιμη όταν πρόκειται για δεδομένα που περιέχουν ακραίες τιμές (στην περίπτωση μας το p('Hallucination')) έχει εύρος τιμών 0 έως 1). Ο συντελεστής συσχέτισης κατάταξης Spearman  $r_s$ , που αναφέρεται επίσης ως συντελεστής κατάταξης, χρησιμοποιείται για την ποσοτικοποίηση αυτής της συσχέτισης. Ο τύπος για το  $r_s$  δίνεται παρακάτω, όπου  $n$  αντιπροσωπεύει τον αριθμό των σημείων δεδομένων στη σειρά και  $d$  υποδηλώνει τη διαφορά στις τάξεις μεταξύ των αντίστοιχων τιμών των δύο μεταβλητών:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Εδώ,  $d_i$  είναι η διαφορά μεταξύ των βαθμών των δύο αντίστοιχων τιμών στα ζεύγη δεδομένων.

### 1.3.3 Baseline System

Εκτός από το σύνολο δεδομένων, οι διοργανωτές του SHROOM παρείχαν επίσης ένα σύστημα αναφοράς. Ως αναφορά, χρησιμοποίησαν ένα LLM [22] για να αξιολογήσουν κατά πόσον οι παραγόμενες υποθέσεις είναι συνεπείς με το παρεχόμενο πλαίσιο. Όπως περιγράφεται στο paper περιγραφής του SHROOM [35] το template του prompt που χρησιμοποιήθηκε προέρχεται από το Self-checkgpt [32] όπως φαίνεται στο 1.3.1. Το σύστημα του [32], το οποίο έχει συγκεντρώσει κάποια προσοχή από την κοινότητα, αποτελεί μια απλή προσέγγιση που βασίζεται σε ένα σύγχρονο LLM, και επομένως είναι κατάλληλο για να χρησιμεύσει ως βασική γραμμή στο SHROOM: αντιστοιχεί σε μια λογική προεπιλεγμένη προσέγγιση για την αντιμετώπιση του προβλήματος που κλήθηκαν να αντιμετωπίσουν οι συμμετέχοντες.

```
Context: {}
Sentence: {}
Is the Sentence supported by the Context above?
Answer using ONLY yes or no:
```

Figure 1.3.1: Prompt template που χρησιμοποιήθηκε στο μοντέλο αναφοράς

Το συγκεκριμένο πλαίσιο ποικίλλει ανάλογα με την εργασία που αντιμετωπίζει, δηλαδή η πρόταση-πηγή για την παραγωγή παραφράσεων και η πρόταση-στόχος για την μηχανική μετάφραση και μοντελοποίηση ορισμών. Όσον αφορά την πιθανότητα hallucination, το σύστημα βασίζεται στην πιθανότητα που αποδίδει το μοντέλο στην πρώτη λέξη εξόδου. Σε περιπτώσεις όπου η έξοδος δεν υποδεικνύει σαφώς *yes* ή *no*, επιλέγεται τυχαία μία από αυτές, αποδίδοντας πιθανότητες hallucination 0.5.

Τα αποτελέσματα αυτής της μεθόδου φαίνονται στον Πίνακα 1.2.

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.65     | 0.38 |
| Val aware     | 0.71     | 0.46 |
| Test agnostic | 0.75     | 0.49 |
| Test aware    | 0.70     | 0.40 |

Table 1.2: Αποτελέσματα του μοντέλου αναφοράς

## 1.4 Μέθοδοι

### Ομοιότητα Προτάσεων (Sentence Similarity)

Η αρχική προσέγγιση για την επίλυση του προβλήματος περιελάμβανε την αναδιατύπωσή του ως πρόβλημα ομοιότητας προτάσεων. Ως πρόβλημα ομοιότητας προτάσεων θεωρούμε την διαδικασία μέτρησης του βαθμού ομοιότητας μεταξύ δύο προτάσεων με βάση το σημασιολογικό τους περιεχόμενο. Η διαδικασία αυτή είναι σημαντική σε πολλές εργασίες επεξεργασίας φυσικής γλώσσας (NLP), καθώς επιτρέπει τη σύγκριση των προτάσεων όχι απλώς σε λεξιλογικό επίπεδο αλλά μέσω του υποκείμενου νοήματός τους. Για την πραγματοποίηση αυτής της εργασίας, τα μοντέλα συνήθως αναπαριστούν τα κείμενα εισόδου ως διανύσματα που κωδικοποιούν το σημασιολογικό τους περιεχόμενο. Η ομοιότητα μεταξύ αυτών των διανυσμάτων στη συνέχεια ποσοτικοποιείται με τον υπολογισμό της γειτνίασης τους εντός ενός διανυσματικού χώρου υψηλής διάστασης.

Στη δική μας υλοποίηση αυτής της μεθόδου, πραγματοποιήσαμε διάφορα βήματα για να προσαρμόσουμε το πρόβλημα μας σε πρόβλημα ομοιότητας προτάσεων. Αρχικά, κωδικοποιήσαμε τόσο το κείμενο-στόχο («tgt») όσο και το κείμενο-υπόθεση («hyp») για κάθε δείγμα δεδομένων. Για να το κάνουμε αυτό, χρησιμοποιήσαμε διάφορα προ-εκπαιδευμένα μοντέλα από τη βιβλιοθήκη Sentence\_transformers, η οποία είναι γνωστή για την παράγωγή υψηλής ποιότητας διανυσμάτων που αποτυπώνουν αποτελεσματικά τις σημασιολογικές αποχρώσεις των προτάσεων. Αυτά τα διανύσματα χρησιμεύουν ως αναπαραστάσεις των προτάσεων, διατηρώντας το νόημά τους, ενώ επιτρέπουν το μαθηματικό τους χειρισμό. Αφού λάβαμε τις ενσωματώσεις, υπολογίσαμε την απόσταση μεταξύ των διανυσμάτων 'tgt' και 'hyp'. Για να εξασφαλίσουμε μια ολοκληρωμένη αξιολόγηση της ομοιότητας, χρησιμοποιήσαμε πολλαπλές μετρικές απόστασης, συγκεκριμένα τη Manhattan απόσταση και την ευκλείδεια

απόσταση. Αυτές οι μετρικές μας επέτρεψαν να ποσοτικοποιήσουμε την έκταση της διαφοράς ή της ομοιότητας μεταξύ των ζευγών προτάσεων, μετρώντας πόσο απέχουν τα αντίστοιχα διανύσματα τους στον διανυσματικό χώρο.

Τέλος, με βάση τις υπολογιζόμενες αποστάσεις, καθορίσαμε ένα κατάλληλο κατώφλι για τη δυαδική ταξινόμηση. Αυτό το κατώφλι καθορίστηκε πειραματικά αφού δοκιμάστηκαν διάφορες τιμές απόστασης, επιτρέποντάς έτσι να ορίσουμε ένα όριο που διακρίνει αποτελεσματικά παρόμοια και ανόμοια ζεύγη προτάσεων. Μέσω αυτής της διαδικασίας, προσαρμόσαμε το πρόβλημά μας ώστε να επιτευχθεί η δυαδική ταξινόμηση. Η μέθοδος αυτή παρείχε ένα αρχικό πλαίσιο για την αντιμετώπιση του προβλήματος, αξιοποιώντας καθιερωμένες τεχνικές στην ομοιότητα προτάσεων

## COMET

Το COMET (Crosslingual Optimized Metric for Evaluation of Translation) είναι ένα προηγμένο νευρωνικό πλαίσιο που αναπτύχθηκε για την αξιολόγηση της μηχανικής μετάφρασης (MT). Αναπτύχθηκε για να αντιμετωπίσει τις ελλείψεις των παραδοσιακών μετρικών αξιολόγησης όπως το BLEU και το METEOR, τα οποία βασίζονται κυρίως στην επικάλυψη λέξεων και αγνοούν τις σημασιολογικές αποχρώσεις. Το COMET χρησιμοποιεί σύγχρονα προεκπαιδευμένα διαγλωσσικά μοντέλα για να αξιολογεί την ποιότητα της μετάφρασης βασιζόμενο στο νόημα και όχι στην ταύτιση λέξεων. Ένα από τα κύρια πλεονεκτήματα του COMET είναι η πολυγλωσσική και διαγλωσσική προσαρμοστικότητα του, που το καθιστά κατάλληλο για πολλά γλωσσικά περιβάλλοντα. Το COMET έχει σχεδιαστεί για να μιμείται την ανθρώπινη κρίση κατά την αξιολόγηση των μεταφράσεων. Εκπαιδεύεται σε ανθρώπινες αξιολογήσεις ποιότητας και προβλέπει την ποιότητα της μετάφρασης με ακρίβεια. Το COMET βασίζεται σε δύο κύριες αρχιτεκτονικές μοντέλων: το μοντέλο εκτιμητή και το μοντέλο κατάταξης της μετάφρασης. Το COMET έχει επιτύχει αξιοσημείωτα αποτελέσματα και ξεχώρισε σε διάφορους διαγωνισμούς.

Το COMET είναι επίσης ευέλικτο και μπορεί να λειτουργήσει τόσο με βάση τις αναφορές όσο και χωρίς αναφορές. Ενώ οι παραδοσιακές μέθοδοι αξιολόγησης της MT απαιτούν αναφορά μετάφρασης που έχει δημιουργηθεί από ανθρώπους, το COMET μπορεί να αξιολογήσει με μόνο το κείμενο προέλευσης. Αυτή η δυνατότητα έχει εμπνευστεί από τα μοντέλα εκτίμησης ποιότητας, τα οποία αξιολογούν την ποιότητα της μετάφρασης χωρίς αναφορά. Αυτή η ευελιξία επιτρέπει την προσαρμογή του COMET σε άλλες εργασίες παραγωγής φυσικής γλώσσας, όπως η αξιολόγηση της συνέπειας και της ρεαλιστικής ακρίβειας των αποτελεσμάτων. Η τεχνική υλοποίηση του COMET περιλαμβάνει έναν νέο μηχανισμό προσοχής ανά στρώμα, που συγκεντρώνει πληροφορίες από διάφορα στρώματα του προ-εκπαιδευμένου μοντέλου κωδικοποιητή. Αυτό ενισχύει την ικανότητα του COMET να κάνει λεπτομερείς προβλέψεις σχετικά με την ποιότητα της μετάφρασης. Το COMET έχει επίσης εξαιρετική απόδοση και έχει ξεχωρίσει σε διάφορους διαγωνισμούς. Είναι ένα εργαλείο αξιολόγησης που είναι διαθέσιμο ως ανοικτός κώδικας, καθιστώντας το προσιτό σε ερευνητές και επαγγελματίες.

## Fine-tune μοντέλα NLI

Στο πλαίσιο της ανίχνευσης hallucination, χρησιμοποιήσαμε επίσης Natural Language Inference(NLI), μια προσέγγιση που έχει σημειώσει σημαντική πρόοδο τα τελευταία χρόνια, ιδίως όσον αφορά την ικανότητά της να συλλαμβάνει και να αναλύει σημασιολογικές πολυπλοκότητες που ευθυγραμμίζονται στενά με το φαινόμενο των hallucination. Τα μοντέλα NLI έχουν σχεδιαστεί για να αξιολογούν τη λογική σχέση μεταξύ δύο προτάσεων, ταξινομώντας τη σύνδεση μεταξύ μιας υπόθεσης και μιας προκειμένης πρότασης σε μία από τρεις κατηγορίες: συνεπαγωγή (entailment), ουδετερότητα (neutral) ή αντίφαση (contradiction). Αυτές οι κατηγορίες αντιστοιχούν στη κατανόηση του μοντέλου για το αν η υπόθεση προκύπτει λογικά από την προκειμένη, είναι άσχετη ή άμεσα αντιφάσκει με αυτήν.

Δεδομένης της ικανότητάς τους να χειρίζονται διαφοροποιημένες σημασιολογικές σχέσεις, τα μοντέλα NLI προσφέρουν ένα πολύτιμο πλαίσιο για αντιμετώπιση του ζητήματος της ανίχνευσης hallucination σε γλωσσικά μοντέλα. Συγκεκριμένα, διαμορφώσαμε το πρόβλημα της ανίχνευσης hallucination ως έργο NLI: δεδομένης της εισόδου που παράγεται από το μοντέλο, την οποία ορίζουμε ως την υπόθεση (συμβολίζεται ως hyp), και τον αντίστοιχο στόχο ή αναφορά (συμβολίζεται ως tgt), ο στόχος είναι να προσδιοριστεί εάν η πρόταση tgt συνεπάγεται, αντιφάσκει ή παραμένει ουδέτερη σε σχέση με την πρόταση hyp. Σε αυτό το πλαίσιο, τα hallucinations θα εκδηλωνόταν ως αντιφάσεις ή ουδέτερες σχέσεις, παρά ως συνεπαγωγές.

Για την εφαρμογή αυτής της προσέγγισης, επιλέξαμε ένα προ-εκπαιδευμένο μοντέλο NLI που είναι διαθέσιμο μέσω του Hugging Face library, συγκεκριμένα με την αρχιτεκτονική mDeBERTa-v3-base1, το οποίο έχει εκ-

παιδεύει σε μια μεγάλης κλίμακας πολύγλωσσο σύνολο δεδομένων. Η χρήση αυτού του μοντέλου προσφέρει σαφή πλεονεκτήματα λόγω της ικανότητάς του να χειρίζεται ένα ευρύ φάσμα γλωσσικών δομών και γλωσσών, καθιστώντας το ιδιαίτερα προσαρμόσιμο στα ποικίλα και πολύπλοκα δεδομένα που συναντήσαμε.

Για να προσαρμόσουμε περαιτέρω το μοντέλο NLI στις ειδικές απαιτήσεις της εργασίας μας, το εκπαιδεύσαμε περαιτέρω (finetuning) χρησιμοποιώντας τα σύνολα επικύρωσης με τις ετικέτες τους. Αυτό το βήμα μας επέτρεψε να διευρύνουμε τις δυνατότητες του μοντέλου στο να εντοπίζει τις λεπτές διακρίσεις που είναι απαραίτητες για τον εντοπισμό hallucination στις παραγόμενες εξόδους. Με αυτό τον τρόπο, ενισχύσαμε την ικανότητά του να διακρίνει αν η παραγόμενη έξοδος από το μοντέλο ευθυγραμμίζεται ή αποκλίνει από την αναφορά, βελτιώνοντας αποτελεσματικά την απόδοσή του σε αυτό το συγκεκριμένο τομέα.

Αυτή η προσέγγιση της μετατροπής της ανίχνευσης hallucination σε έργο NLI δεν παρείχε μόνο ένα σαφές πλαίσιο για την αξιολόγηση της λογικής συνοχής μεταξύ της παραγόμενης εξόδου και της αναφοράς, αλλά μας επέτρεψε επίσης να αξιοποιήσουμε τα πλεονεκτήματα των σύγχρονων μοντέλων NLI. Με την εφαρμογή αυτών των μοντέλων, μπορέσαμε να φέρουμε μια δομημένη, τυπική μέθοδο στο έργο της ανίχνευσης hallucination, προσφέροντας βελτιωμένη ακρίβεια στον εντοπισμό τότε το παραγόμενο κείμενο ήταν συνεπές με την αναμενόμενη έξοδο ή παρουσίαζε στοιχεία ψευδαίσθησης.

### **Fine-tune προ-εκπαιδευμένου μοντέλου για ανίχνευση hallucination**

Η ανίχνευση hallucination είναι μια σημαντική πρόκληση στον τομέα της επεξεργασίας φυσικής γλώσσας. Σε αυτό το κείμενο παρουσιάζεται μια προσέγγιση που χρησιμοποιεί ένα προ-εκπαιδευμένο μοντέλο για την ανίχνευση hallucination στο σύνολο δεδομένων SHROOM. Μετά από την αρχική εκπαίδευση στα δεδομένα συνεπαγωγής κειμένου, το μοντέλο υποβάλλεται σε περαιτέρω εκπαίδευση σε σύνολα δεδομένων περίληψης, προκειμένου να βελτιωθεί η ικανότητά του να ανιχνεύει τα hallucinations. Το τελικό μοντέλο παράγει μια πιθανολογική έξοδο που δίνει μια βαθμολογία πιθανότητας για το κατά πόσο ένα κείμενο περιλαμβάνει hallucination.

Η διαδικασία εκπαίδευσης του μοντέλου γίνεται με τη χρήση επικύρωσης, όπου το μοντέλο εκπαιδεύεται σε ένα μικρό υποσύνολο του συνόλου δεδομένων SHROOM. Μετά το πέρας της εκπαίδευσης, το μοντέλο αξιολογείται χρησιμοποιώντας ένα κατώφλι, το οποίο καθορίζει ποια εξόδος θεωρείται ως ορθή και ποια ως hallucination. Η εφαρμογή αυτής της προσέγγισης επιτρέπει την αξιολόγηση των αποτελεσμάτων και τη λήψη αποφάσεων για την παρουσία hallucination.

Η προτεινόμενη προσέγγιση έχει αρκετά πλεονεκτήματα. Καταρχάς, το προ-εκπαιδευμένο μοντέλο που χρησιμοποιείται έχει εκπαιδευτεί σε δεδομένα συνεπαγωγής κειμένου, προσδίδοντας ισχυρή ικανότητα στο μοντέλο να αναγνωρίζει τη συνοχή και τις σχέσεις μεταξύ των προτάσεων. Δεύτερον, η προσέγγιση επιτρέπει την ευέλικτη κατωφλίωση, δίνοντας τη δυνατότητα προσαρμογής του κατωφλίου ανάλογα με το επίπεδο ανοχής για hallucinations. Τέλος, ο μηχανισμός κατωφλίου παρέχει μια δυαδική ταξινόμηση για παρουσία hallucination, καθιστώντας την απόφαση πιο απλή για πρακτική χρήση.

Όλα αυτά τα χαρακτηριστικά καθιστούν την προτεινόμενη προσέγγιση αποτελεσματική και αξιόπιστη στον τομέα της ανίχνευσης hallucination. Οι πειραματικές αξιολογήσεις έχουν δείξει ότι το μοντέλο είναι ευαίσθητο στα hallucinations και μπορεί να αναγνωρίσει με ακρίβεια τις περιπτώσεις που αποκλίνουν από την πραγματολογική ακρίβεια. Μελλοντικές επεκτάσεις αυτής της προσεγγίσεως μπορούν να περιλαμβάνουν τον εκσυγχρονισμό των δεδομένων εκπαίδευσης με ειδικά σενάρια hallucination και τη χρήση πιο πολύπλοκων μοντέλων που θα βελτιώσουν περαιτέρω την ακρίβεια και την ευελιξία του συστήματος ανίχνευσης.

### **Ταξινομητής Ψηφοφορίας**

Η τελική μας μεθοδολογική προσέγγιση βασίζεται στην χρήση ενός ταξινομητή ψηφοφορίας για τη συνένωση των προβλέψεων από πολλαπλά μοντέλα. Ο ταξινομητής ψηφοφορίας συγχωνεύει τις εισόδους των μοντέλων και παράγει ένα πιο αξιόπιστο και ακριβές αποτέλεσμα βασιζόμενος στην πλειοψηφία των προβλέψεων. Αυτή η προσέγγιση επιτρέπει στον ταξινομητή ψηφοφορίας να αξιοποιεί τα μοναδικά πλεονεκτήματα κάθε μοντέλου και να διορθώνει τυχόν αδυναμίες ή περιορισμούς ενός μεμονωμένου μοντέλου.

Ο ταξινομητής ψηφοφορίας μπορεί να ανιχνεύει διάφορους τύπους αποκλίσεων, όπως τις συντακτικές και σημασιολογικές ασυνέπειες, επωφελούμενος από τις διαφορετικές δυνατότητες κάθε μοντέλου. Η συγχέντρωση αυτών

των διαφορετικών προοπτικών οδηγεί σε πιο ακριβείς και ολοκληρωμένες προβλέψεις, ενισχύοντας τη συνολική απόδοση και καθιστώντας το σύστημα πιο ανθεκτικό στην πολυπλοκότητα των δεδομένων.

Επιπλέον, ο ταξινομητής ψηφοφορίας παρέχει μεγαλύτερη ευελιξία στις αποφάσεις του, καθώς μπορεί να υπολογίζει το ποσοστό των μοντέλων που ταξινομούν ένα αποτέλεσμα ως hallucination. Αυτή η προσέγγιση παρέχει μια πιο λεπτομερή εικόνα του βαθμού hallucination σε ένα σημείο δεδομένων και επιτρέπει την ιεράρχηση των περιπτώσεων για περαιτέρω ανάλυση ή βελτίωση.

Η προσέγγιση του συνόλου είναι ιδιαίτερα χρήσιμη στην ανίχνευση hallucination, καθώς βοηθά στην αντιμετώπιση της πολυπλοκότητας και της μεταβλητότητας των δεδομένων. Με τη συνδυασμένη προβολή των προβλέψεων πολλαπλών μοντέλων, ο ταξινομητής ψηφοφορίας μπορεί να πετύχει μια πιο ακριβή κατανόηση των hallucination. Με αυτήν την προσέγγιση, δημιουργήσαμε ένα ισχυρό και προσαρμόσιμο σύστημα για την ανίχνευση hallucination.

Συνολικά, ο ταξινομητής ψηφοφορίας δεν βελτιώνει μόνο την ακρίβεια των προβλέψεων, αλλά προσφέρει και μια πιο διαφοροποιημένη προσέγγιση για την κατανόηση των πολύπλοκων αποτελεσμάτων και την αξιολόγηση της εμπιστοσύνης στις προβλέψεις. Η προσέγγιση του συνόλου κερδίζει τις προκλήσεις που εμφανίζονται στην ανίχνευση hallucination και επιβεβαιώνει την αξιοπιστία και την επεκτασιμότητα της λύσης μας.

## Prompting

Στην τελική μας προσέγγιση για την ανίχνευση hallucination, εφαρμόσαμε τη μέθοδο Prompting των Μεγάλων Γλωσσικών Μοντέλων (LLM) ως μια εξελιγμένη στρατηγική για την ενίσχυση της ακρίβειας και της αξιοπιστίας των παραγόμενων αποτελεσμάτων. Αυτή η μέθοδος αξιοποιεί τις προηγμένες δυνατότητες των τελευταίας τεχνολογίας LLM, επιτρέποντάς μας να παράγουμε, να αξιολογούμε και να βελτιώνουμε συστηματικά τις εξόδους για τον εντοπισμό και τον μετριασμό του περιεχομένου hallucination.

Για να επιτύχουμε τους στόχους μας, χρησιμοποιήσαμε ένα ευρύ φάσμα από LLMs, συμπεριλαμβανομένων αξιολογούμενων μοντέλων όπως το Gemma, το Mistral και το Llama. Αυτά τα μοντέλα επιλέχθηκαν για τις ξεχωριστές αρχιτεκτονικές και τα δυνατά τους σημεία, τα οποία μας παρείχαν μια ολοκληρωμένη εργαλειοθήκη για την ανίχνευση hallucination. Η μεθοδολογία μας ενσωμάτωσε μια ποικιλία τεχνικών προτροπής που αποσκοπούσαν στη μεγιστοποίηση των δυνατοτήτων των μοντέλων για τον ακριβή εντοπισμό και την αντιμετώπιση hallucination. Συγκεκριμένα, χρησιμοποιήσαμε τεχνικές με zero-shot prompts, chain of thought reasoning, και few-shot.

Η τεχνική zero-shot μας επέτρεψε να ζητήσουμε από τα μοντέλα να εκτελέσουν εργασίες χωρίς προηγούμενα παραδείγματα, επιτρέποντάς τους να βασίζονται στην εκπαιδευμένη γνώση τους για να παράγουν απαντήσεις. Αυτή η προσέγγιση ήταν ιδιαίτερα χρήσιμη σε σενάρια όπου τα συγκεκριμένα παραδείγματα hallucination ήταν ελάχιστα, επιτρέποντας ευρύτερη γενίκευση στα αποτελέσματα των μοντέλων.

## 1.5 Πειραματικό Μέρος

### 1.5.1 Ομοιότητα προτάσεων

Ξεκινώντας τα πειράματά μας, χρησιμοποιήσαμε διάφορα προ-εκπαιδευμένα μοντέλα αρχιτεκτονικής transformer για να κωδικοποιήσουμε το «tgt» και το «hyp» κάθε δείγματος των δεδομένων στο validation σύνολο δεδομένων, τόσο στο σύνολο δεδομένων model-aware όσο και στο σύνολο δεδομένων model-agnostic. Το επόμενο βήμα ήταν ο υπολογισμός του σκορ ομοιότητας αυτών των δύο προτάσεων. Για να το κάνουμε αυτό, χρησιμοποιήσαμε διάφορες μεθόδους υπολογισμού αποστάσεων, όπως η ευκλείδεια απόσταση, η απόσταση manhattan, η ομοιότητα συνημίτονου και μεθόδους clustering.

#### Υπολογισμός απόστασης

Για τις υπολογιζόμενες αποστάσεις, ήταν απαραίτητο να καθοριστεί ένα κατάλληλο κατώφλι για την εκτέλεση δυαδικής ταξινόμησης, όπου κάθε σημείο δεδομένων θα χαρακτηριζόταν είτε ως «Hallucination» είτε ως «Not Hallucination». Μετά τη διεξαγωγή μιας σειράς διερευνητικών πειραμάτων για τον προσδιορισμό ενός κατάλληλου κατωφλίου, παρατηρήσαμε ότι επιτυγχάνεται μεγαλύτερη ακρίβεια ταξινόμησης όταν χρησιμοποιείται

ως κατώφλι η διάμεσος των τιμών των αποστάσεων. Η διάμεσος διαχωρίζει αποτελεσματικά το σύνολο δεδομένων σε δύο μισά, διακρίνοντας τις υψηλότερες τιμές από τις χαμηλότερες. Τα κατώτατα όρια που προέκυψαν για κάθε μοντέλο και σύνολο δεδομένων παρουσιάζονται στο Appendix A.

Αφού καθορίστηκαν τα κατώφλια, η δυαδική ταξινόμηση πραγματοποιήθηκε σύμφωνα με την ακόλουθη διαδικασία:

- **Για τις ευκλείδειες αποστάσεις και τις αποστάσεις Manhattan**, αρχικά κανονικοποιήσαμε τις υπολογισμένες αποστάσεις για να περιορίσουμε τις τιμές τους εντός του εύρους  $[0, 1]$ . Μετά την κανονικοποίηση, ένα σημείο δεδομένων ταξινομήθηκε ως «Hallucination» εάν η κανονικοποιημένη απόστασή του υπερέβαινε το κατώφλι. Στην περίπτωση αυτή, η πιθανότητα hallucination, που συμβολίζεται ως  $p(\text{'Hallucination'})$ , ορίστηκε ίση με την αντίστοιχη κανονικοποιημένη απόσταση.
- **Για την ομοιότητα συνημίτονου**, η οποία εκ φύσεως αποδίδει μια βαθμολογία ομοιότητας μεταξύ 0 και 1, συγκρίναμε τη βαθμολογία με το κατώφλι. Εάν η βαθμολογία ομοιότητας ήταν χαμηλότερη από το κατώφλι, το σημείο δεδομένων χαρακτηρίστηκε ως «Hallucination». Σε αυτό το σενάριο, η πιθανότητα hallucination,  $p(\text{'Hallucination'})$ , υπολογίστηκε ως ένα μείον τη βαθμολογία ομοιότητας (δηλαδή,  $p(\text{'Hallucination'}) = 1 - \text{βαθμολογία ομοιότητας}$ ).

Τα αποτελέσματα για τη μετρική ακρίβειας παρουσιάζονται στους πίνακες 1.3,1.4,1.5,1.6. Τα αποτελέσματα για τη μετρική συσχέτισης Spearman παρέχονται στους πίνακες 1.7,1.8,1.9,1.10

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.61             | 0.59                  | 0.61             |
| stsb-roberta-large                                                   | 0.7              | 0.7                   | 0.7              |
| all-distilroberta-v1                                                 | 0.65             | 0.65                  | 0.65             |
| all-MiniLM-L6-v2                                                     | 0.66             | 0.66                  | 0.66             |
| all-MiniLM-L12-v2                                                    | 0.66             | 0.66                  | 0.66             |
| all-mpnet-base-v2                                                    | 0.67             | 0.67                  | 0.67             |
| all-roberta-large-v1                                                 | 0.65             | 0.65                  | 0.65             |
| paraphrase-MiniLM-L6-v2                                              | 0.66             | 0.67                  | 0.67             |
| paraphrase-mpnet-base-v2                                             | 0.7              | 0.7                   | 0.7              |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.66             | 0.66                  | 0.66             |
| paraphrase-MiniLM-L12                                                | 0.67             | 0.67                  | 0.67             |
| paraphrase-MiniLM-L3                                                 | 0.65             | 0.66                  | 0.65             |
| paraphrase-TinyBERT-L6                                               | 0.68             | 0.68                  | 0.68             |
| paraphrase-distilroberta-base                                        | 0.66             | 0.69                  | 0.67             |
| xlm-r-distilroberta                                                  | 0.67             | 0.68                  | 0.67             |
| stsb-roberta-base-v2                                                 | 0.69             | 0.7                   | 0.69             |
| stsb-roberta-large                                                   | 0.7              | 0.7                   | 0.7              |
| stsb-distilroberta-base-v2                                           | 0.67             | 0.67                  | 0.68             |
| stsb-mpnet-base-v2                                                   | 0.7              | 0.7                   | 0.7              |
| multi-qa-distilbert-cos-v1                                           | 0.66             | 0.66                  | 0.66             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.66             | 0.66                  | 0.66             |
| sentence-t5-base                                                     | 0.71             | 0.71                  | 0.71             |
| sentence-t5-large                                                    | <b>0.72</b>      | <b>0.72</b>           | <b>0.72</b>      |
| gtr-t5-base                                                          | 0.68             | 0.68                  | 0.68             |
| gtr-t5-large                                                         | 0.68             | 0.68                  | 0.68             |
| nq-distilbert-base-v1                                                | 0.65             | 0.65                  | 0.65             |
| nli-distilroberta-base-v2                                            | 0.68             | 0.69                  | 0.68             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.68             | 0.67                  | 0.67             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.69             | 0.7                   | 0.69             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.63             | 0.63                  | 0.64             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.67             | 0.68                  | 0.67             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.68             | 0.67                  | 0.67             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.68             | 0.67                  | 0.67             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.68             | 0.68                  | 0.67             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.63             | 0.63                  | 0.64             |

Table 1.3: Accuracy του test model-agnostic συνόλου δεδομένων

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.55             | 0.55                  | 0.54             |
| stsb-roberta-large                                                   | 0.59             | 0.6                   | 0.6              |
| all-distilroberta-v1                                                 | 0.58             | 0.58                  | 0.59             |
| all-MiniLM-L6-v2                                                     | 0.57             | 0.57                  | 0.57             |
| all-MiniLM-L12-v2                                                    | 0.58             | 0.58                  | 0.58             |
| all-mpnet-base-v2                                                    | 0.58             | 0.58                  | 0.58             |
| all-roberta-large-v1                                                 | 0.59             | 0.59                  | 0.58             |
| paraphrase-MiniLM-L6-v2                                              | 0.56             | 0.58                  | 0.56             |
| paraphrase-mpnet-base-v2                                             | 0.59             | 0.59                  | 0.59             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.57             | 0.57                  | 0.56             |
| paraphrase-MiniLM-L12                                                | 0.56             | 0.57                  | 0.56             |
| paraphrase-MiniLM-L3                                                 | 0.55             | 0.57                  | 0.56             |
| paraphrase-TinyBERT-L6                                               | 0.56             | 0.58                  | 0.56             |
| paraphrase-distilroberta-base                                        | 0.57             | 0.59                  | 0.57             |
| xlm-r-distilroberta                                                  | 0.59             | 0.58                  | 0.58             |
| stsb-roberta-base-v2                                                 | 0.57             | 0.58                  | 0.57             |
| stsb-roberta-large                                                   | 0.59             | 0.6                   | 0.6              |
| stsb-distilroberta-base-v2                                           | 0.57             | 0.57                  | 0.57             |
| stsb-mpnet-base-v2                                                   | 0.58             | 0.59                  | 0.58             |
| multi-qa-distilbert-cos-v1                                           | 0.57             | 0.57                  | 0.57             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.57             | 0.57                  | 0.56             |
| sentence-t5-base                                                     | 0.6              | 0.6                   | 0.6              |
| sentence-t5-large                                                    | <b>0.62</b>      | <b>0.62</b>           | <b>0.62</b>      |
| gtr-t5-base                                                          | 0.58             | 0.58                  | 0.58             |
| gtr-t5-large                                                         | 0.59             | 0.59                  | 0.59             |
| nq-distilbert-base-v1                                                | 0.58             | 0.58                  | 0.58             |
| nli-distilroberta-base-v2                                            | 0.58             | 0.58                  | 0.58             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.57             | 0.58                  | 0.57             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.58             | 0.58                  | 0.58             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.56             | 0.56                  | 0.56             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.59             | 0.58                  | 0.58             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.57             | 0.57                  | 0.57             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.57             | 0.57                  | 0.57             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.58             | 0.59                  | 0.58             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.56             | 0.56                  | 0.56             |

Table 1.4: Accuracy του test model-aware συνόλου δεδομένων

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.67             | 0.65                  | 0.67             |
| stsb-roberta-large                                                   | 0.74             | 0.74                  | 0.74             |
| all-distilroberta-v1                                                 | 0.68             | 0.68                  | 0.68             |
| all-MiniLM-L6-v2                                                     | 0.69             | 0.69                  | 0.69             |
| all-MiniLM-L12-v2                                                    | 0.69             | 0.69                  | 0.69             |
| all-mpnet-base-v2                                                    | 0.68             | 0.68                  | 0.68             |
| all-roberta-large-v1                                                 | 0.71             | 0.71                  | 0.71             |
| paraphrase-MiniLM-L6-v2                                              | 0.74             | 0.71                  | 0.74             |
| paraphrase-mpnet-base-v2                                             | 0.72             | 0.72                  | 0.72             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.71             | 0.71                  | 0.71             |
| paraphrase-MiniLM-L12                                                | 0.74             | 0.72                  | 0.74             |
| paraphrase-MiniLM-L3                                                 | 0.7              | 0.7                   | 0.69             |
| paraphrase-TinyBERT-L6                                               | 0.71             | 0.72                  | 0.7              |
| paraphrase-distilroberta-base                                        | 0.72             | 0.73                  | 0.72             |
| xlm-r-distilroberta                                                  | 0.72             | 0.73                  | 0.72             |
| stsb-roberta-base-v2                                                 | 0.75             | 0.75                  | 0.75             |
| stsb-roberta-large                                                   | 0.74             | 0.74                  | 0.74             |
| stsb-distilroberta-base-v2                                           | 0.73             | 0.74                  | 0.72             |
| stsb-mpnet-base-v2                                                   | <b>0.75</b>      | <b>0.75</b>           | <b>0.75</b>      |
| multi-qa-distilbert-cos-v1                                           | 0.69             | 0.69                  | 0.69             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.71             | 0.71                  | 0.71             |
| sentence-t5-base                                                     | 0.73             | 0.73                  | 0.73             |
| sentence-t5-large                                                    | 0.74             | 0.74                  | 0.74             |
| gtr-t5-base                                                          | 0.69             | 0.69                  | 0.69             |
| gtr-t5-large                                                         | 0.7              | 0.7                   | 0.7              |
| nq-distilbert-base-v1                                                | 0.69             | 0.69                  | 0.69             |
| nli-distilroberta-base-v2                                            | 0.72             | 0.72                  | 0.72             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.73             | 0.72                  | 0.73             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.73             | 0.73                  | 0.74             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.66             | 0.66                  | 0.67             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.72             | 0.73                  | 0.72             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.72             | 0.72                  | 0.72             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.72             | 0.72                  | 0.72             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.72             | 0.73                  | 0.73             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.66             | 0.66                  | 0.67             |

Table 1.5: Accuracy του validation model-agnostic συνόλου δεδομένων

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.53             | 0.53                  | 0.53             |
| stsb-roberta-large                                                   | 0.58             | 0.59                  | 0.58             |
| all-distilroberta-v1                                                 | 0.57             | 0.57                  | 0.56             |
| all-MiniLM-L6-v2                                                     | 0.56             | 0.56                  | 0.56             |
| all-MiniLM-L12-v2                                                    | 0.57             | 0.57                  | 0.57             |
| all-mpnet-base-v2                                                    | 0.56             | 0.56                  | 0.56             |
| all-roberta-large-v1                                                 | 0.57             | 0.57                  | 0.57             |
| paraphrase-MiniLM-L6-v2                                              | 0.55             | 0.57                  | 0.55             |
| paraphrase-mpnet-base-v2                                             | 0.57             | 0.59                  | 0.57             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.55             | 0.55                  | 0.55             |
| paraphrase-MiniLM-L12                                                | 0.57             | 0.57                  | 0.57             |
| paraphrase-MiniLM-L3                                                 | 0.54             | 0.56                  | 0.54             |
| paraphrase-TinyBERT-L6                                               | 0.55             | 0.56                  | 0.55             |
| paraphrase-distilroberta-base                                        | 0.55             | 0.55                  | 0.55             |
| xlm-r-distilroberta                                                  | 0.57             | 0.57                  | 0.56             |
| stsb-roberta-base-v2                                                 | 0.56             | 0.58                  | 0.56             |
| stsb-roberta-large                                                   | 0.58             | 0.59                  | 0.58             |
| stsb-distilroberta-base-v2                                           | 0.55             | 0.56                  | 0.56             |
| stsb-mpnet-base-v2                                                   | 0.56             | 0.56                  | 0.57             |
| multi-qa-distilbert-cos-v1                                           | 0.55             | 0.55                  | 0.55             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.55             | 0.55                  | 0.55             |
| sentence-t5-base                                                     | 0.58             | 0.58                  | 0.58             |
| sentence-t5-large                                                    | <b>0.61</b>      | <b>0.61</b>           | <b>0.6</b>       |
| gtr-t5-base                                                          | 0.56             | 0.56                  | 0.55             |
| gtr-t5-large                                                         | 0.57             | 0.57                  | 0.57             |
| nq-distilbert-base-v1                                                | 0.56             | 0.57                  | 0.56             |
| nli-distilroberta-base-v2                                            | 0.57             | 0.57                  | 0.57             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.55             | 0.58                  | 0.55             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.58             | 0.59                  | 0.57             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.53             | 0.53                  | 0.54             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.57             | 0.57                  | 0.56             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.55             | 0.55                  | 0.55             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.55             | 0.55                  | 0.55             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.57             | 0.58                  | 0.57             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.53             | 0.53                  | 0.54             |

Table 1.6: Accuracy του validation model-aware συνόλου δεδομένων

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.35             | 0.33                  | 0.36             |
| stsb-roberta-large                                                   | 0.56             | 0.56                  | 0.56             |
| all-distilroberta-v1                                                 | 0.43             | 0.43                  | 0.43             |
| all-MiniLM-L6-v2                                                     | 0.43             | 0.43                  | 0.43             |
| all-MiniLM-L12-v2                                                    | 0.44             | 0.44                  | 0.44             |
| all-mpnet-base-v2                                                    | 0.47             | 0.47                  | 0.47             |
| all-roberta-large-v1                                                 | 0.45             | 0.45                  | 0.45             |
| paraphrase-MiniLM-L6-v2                                              | 0.46             | 0.47                  | 0.46             |
| paraphrase-mpnet-base-v2                                             | 0.54             | 0.53                  | 0.54             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.46             | 0.46                  | 0.46             |
| paraphrase-MiniLM-L12                                                | 0.5              | 0.49                  | 0.51             |
| paraphrase-MiniLM-L3                                                 | 0.41             | 0.43                  | 0.41             |
| paraphrase-TinyBERT-L6                                               | 0.48             | 0.5                   | 0.48             |
| paraphrase-distilroberta-base                                        | 0.48             | 0.5                   | 0.48             |
| xlm-r-distilroberta                                                  | 0.48             | 0.49                  | 0.48             |
| stsb-roberta-base-v2                                                 | 0.54             | 0.56                  | 0.54             |
| stsb-roberta-large                                                   | 0.56             | 0.56                  | 0.56             |
| stsb-distilroberta-base-v2                                           | 0.49             | 0.51                  | 0.49             |
| stsb-mpnet-base-v2                                                   | 0.56             | 0.56                  | 0.56             |
| multi-qa-distilbert-cos-v1                                           | 0.46             | 0.46                  | 0.46             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.46             | 0.46                  | 0.46             |
| sentence-t5-base                                                     | 0.57             | 0.57                  | 0.57             |
| sentence-t5-large                                                    | 0.59             | 0.59                  | 0.59             |
| gtr-t5-base                                                          | 0.48             | 0.48                  | 0.48             |
| gtr-t5-large                                                         | 0.5              | 0.5                   | 0.5              |
| nq-distilbert-base-v1                                                | 0.42             | 0.42                  | 0.42             |
| nli-distilroberta-base-v2                                            | 0.49             | 0.52                  | 0.49             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.5              | 0.49                  | 0.5              |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.54             | 0.55                  | 0.54             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.41             | 0.41                  | 0.41             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.48             | 0.49                  | 0.48             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.49             | 0.49                  | 0.49             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.49             | 0.49                  | 0.49             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.5              | 0.51                  | 0.5              |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.41             | 0.41                  | 0.41             |

Table 1.7: Συσχέτιση Spearman του test model-agnostic συνόλου δεδομένων

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.04             | 0.04                  | 0.05             |
| stsb-roberta-large                                                   | 0.11             | 0.12                  | 0.2              |
| all-distilroberta-v1                                                 | 0.12             | 0.12                  | 0.14             |
| all-MiniLM-L6-v2                                                     | 0.1              | 0.1                   | 0.11             |
| all-MiniLM-L12-v2                                                    | 0.11             | 0.11                  | 0.14             |
| all-mpnet-base-v2                                                    | 0.1              | 0.1                   | 0.1              |
| all-roberta-large-v1                                                 | 0.12             | 0.12                  | 0.12             |
| paraphrase-MiniLM-L6-v2                                              | 0.12             | 0.11                  | 0.16             |
| paraphrase-mpnet-base-v2                                             | 0.11             | 0.12                  | 0.11             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.12             |
| paraphrase-MiniLM-L12                                                | 0.09             | 0.11                  | 0.11             |
| paraphrase-MiniLM-L3                                                 | 0.06             | 0.11                  | 0.07             |
| paraphrase-TinyBERT-L6                                               | 0.14             | 0.12                  | 0.16             |
| paraphrase-distilroberta-base                                        | 0.13             | 0.13                  | 0.14             |
| xlm-r-distilroberta                                                  | 0.1              | 0.1                   | 0.1              |
| stsb-roberta-base-v2                                                 | 0.16             | 0.13                  | 0.2              |
| stsb-roberta-large                                                   | 0.11             | 0.12                  | 0.2              |
| stsb-distilroberta-base-v2                                           | 0.11             | 0.11                  | 0.16             |
| stsb-mpnet-base-v2                                                   | 0.11             | 0.14                  | 0.15             |
| multi-qa-distilbert-cos-v1                                           | 0.1              | 0.1                   | 0.11             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.12             |
| sentence-t5-base                                                     | 0.15             | 0.15                  | 0.15             |
| sentence-t5-large                                                    | 0.16             | 0.16                  | 0.16             |
| gtr-t5-base                                                          | 0.15             | 0.15                  | 0.15             |
| gtr-t5-large                                                         | 0.14             | 0.14                  | 0.14             |
| nq-distilbert-base-v1                                                | 0.15             | 0.11                  | 0.18             |
| nli-distilroberta-base-v2                                            | 0.1              | 0.11                  | 0.13             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.15             | 0.11                  | 0.2              |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.12             | 0.13                  | 0.12             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.12             | 0.06                  | 0.13             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.1              | 0.1                   | 0.1              |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.14             | 0.13                  | 0.14             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.14             | 0.13                  | 0.14             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.15             | 0.14                  | 0.16             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.12             | 0.06                  | 0.13             |

Table 1.8: Συσχέτιση Spearman του test model-aware συνόλου δεδομένων

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.44             | 0.43                  | 0.44             |
| stsb-roberta-large                                                   | 0.63             | 0.63                  | 0.63             |
| all-distilroberta-v1                                                 | 0.53             | 0.53                  | 0.52             |
| all-MiniLM-L6-v2                                                     | 0.54             | 0.54                  | 0.54             |
| all-MiniLM-L12-v2                                                    | 0.53             | 0.53                  | 0.53             |
| all-mpnet-base-v2                                                    | 0.55             | 0.55                  | 0.54             |
| all-roberta-large-v1                                                 | 0.54             | 0.54                  | 0.54             |
| paraphrase-MiniLM-L6-v2                                              | 0.57             | 0.57                  | 0.57             |
| paraphrase-mpnet-base-v2                                             | 0.62             | 0.62                  | 0.61             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.56             | 0.56                  | 0.56             |
| paraphrase-MiniLM-L12                                                | 0.61             | 0.59                  | 0.61             |
| paraphrase-MiniLM-L3                                                 | 0.52             | 0.53                  | 0.52             |
| paraphrase-TinyBERT-L6                                               | 0.56             | 0.59                  | 0.56             |
| paraphrase-distilroberta-base                                        | 0.6              | 0.62                  | 0.6              |
| xlm-r-distilroberta                                                  | 0.57             | 0.59                  | 0.57             |
| stsb-roberta-base-v2                                                 | 0.63             | 0.63                  | 0.63             |
| stsb-roberta-large                                                   | 0.63             | 0.63                  | 0.63             |
| stsb-distilroberta-base-v2                                           | 0.6              | 0.62                  | 0.6              |
| stsb-mpnet-base-v2                                                   | 0.64             | 0.64                  | 0.64             |
| multi-qa-distilbert-cos-v1                                           | 0.56             | 0.56                  | 0.56             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.56             | 0.56                  | 0.56             |
| sentence-t5-base                                                     | 0.63             | 0.63                  | 0.63             |
| sentence-t5-large                                                    | 0.65             | 0.65                  | 0.65             |
| gtr-t5-base                                                          | 0.56             | 0.56                  | 0.56             |
| gtr-t5-large                                                         | 0.58             | 0.58                  | 0.58             |
| nq-distilbert-base-v1                                                | 0.53             | 0.53                  | 0.53             |
| nli-distilroberta-base-v2                                            | 0.6              | 0.63                  | 0.6              |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.62             | 0.6                   | 0.62             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.64             | 0.63                  | 0.63             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.54             | 0.53                  | 0.54             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.57             | 0.59                  | 0.57             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.58             | 0.58                  | 0.58             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.58             | 0.58                  | 0.58             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.62             | 0.62                  | 0.62             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.54             | 0.53                  | 0.54             |

Table 1.9: Συσχέτιση Spearman του model-agnostic συνόλου δεδομένων

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.05             | 0.06                  | 0.07             |
| stsb-roberta-large                                                   | 0.13             | 0.13                  | 0.23             |
| all-distilroberta-v1                                                 | 0.12             | 0.12                  | 0.15             |
| all-MiniLM-L6-v2                                                     | 0.1              | 0.1                   | 0.12             |
| all-MiniLM-L12-v2                                                    | 0.11             | 0.11                  | 0.16             |
| all-mpnet-base-v2                                                    | 0.11             | 0.11                  | 0.11             |
| all-roberta-large-v1                                                 | 0.12             | 0.12                  | 0.12             |
| paraphrase-MiniLM-L6-v2                                              | 0.14             | 0.13                  | 0.18             |
| paraphrase-mpnet-base-v2                                             | 0.13             | 0.14                  | 0.13             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.13             |
| paraphrase-MiniLM-L12                                                | 0.1              | 0.12                  | 0.13             |
| paraphrase-MiniLM-L3                                                 | 0.07             | 0.12                  | 0.07             |
| paraphrase-TinyBERT-L6                                               | 0.15             | 0.13                  | 0.17             |
| paraphrase-distilroberta-base                                        | 0.15             | 0.16                  | 0.16             |
| xlm-r-distilroberta                                                  | 0.11             | 0.12                  | 0.11             |
| stsb-roberta-base-v2                                                 | 0.17             | 0.14                  | 0.21             |
| stsb-roberta-large                                                   | 0.13             | 0.13                  | 0.23             |
| stsb-distilroberta-base-v2                                           | 0.16             | 0.15                  | 0.2              |
| stsb-mpnet-base-v2                                                   | 0.13             | 0.15                  | 0.15             |
| multi-qa-distilbert-cos-v1                                           | 0.1              | 0.1                   | 0.11             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.13             |
| sentence-t5-base                                                     | 0.17             | 0.17                  | 0.17             |
| sentence-t5-large                                                    | 0.19             | 0.19                  | 0.19             |
| gtr-t5-base                                                          | 0.16             | 0.16                  | 0.16             |
| gtr-t5-large                                                         | 0.15             | 0.15                  | 0.16             |
| nq-distilbert-base-v1                                                | 0.17             | 0.12                  | 0.19             |
| nli-distilroberta-base-v2                                            | 0.13             | 0.15                  | 0.17             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.16             | 0.13                  | 0.21             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.14             | 0.15                  | 0.14             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.15             | 0.08                  | 0.16             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.11             | 0.12                  | 0.11             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.15             | 0.14                  | 0.15             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.15             | 0.14                  | 0.15             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.16             | 0.16                  | 0.17             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.15             | 0.08                  | 0.16             |

Table 1.10: Συσχέτιση Spearman του validation model-aware συνόλου δεδομένων

## Ομαδοποίηση

Μια άλλη μέθοδος που χρησιμοποιήσαμε για να βελτιώσουμε τη διαδικασία ταξινόμησης ήταν η ομαδοποίηση. Συγκεκριμένα, υπολογίσαμε την απόσταση μεταξύ του στόχου (tgt) και της υπόθεσης (hyp) προσδιορίζοντας την απόλυτη διαφορά μεταξύ των αντίστοιχων διανυσματικών αναπαραστάσεων τους. Αυτή η μετρική χρησίμευσε ως βάση για την ομαδοποίηση των σημείων δεδομένων. Για τον σκοπό αυτό, χρησιμοποιήσαμε τον αλγόριθμο K-means, ο οποίος είναι κατάλληλος για την κατάτμηση δεδομένων σε διακριτές συστάδες με βάση την ομοιότητα. Θέτοντας τον αριθμό των συστάδων σε δύο, επιδιώξαμε να κατηγοριοποιήσουμε τα ζεύγη στόχων-υποθέσεων σε δύο ομάδες, εντοπίζοντας έτσι μοτίβα ή κοινά σημεία μεταξύ των δεδομένων.

Τα αποτελέσματα της συσταδοποίησης ήταν συνεπή με τα ευρήματα που προέκυψαν μέσω της μεθόδου ταξινόμησης με βάση την απόσταση που περιγράφηκε στο προηγούμενο υποκεφάλαιο. Όπως φαίνεται στον πίνακα 1.11, η προσέγγιση ομαδοποίησης παρείχε παρόμοιες κατηγοριοποιήσεις, ενισχύοντας περαιτέρω την ανθεκτικότητα της μεθόδου υπολογισμού απόστασης. Αυτά τα συγκρίσιμα αποτελέσματα υποδηλώνουν ότι και οι δύο τεχνικές είναι έγκυρες και συμπληρωματικές για το συγκεκριμένο έργο, προσφέροντας διαφορετικές οπτικές γωνίες για την υποκείμενη κατανομή των δεδομένων, ενώ παράλληλα αποδίδουν σύμφωνες ταξινομήσεις.

| Model                                                                 | Test-agnostic | Test-aware | Val-agnostic | Val-aware |
|-----------------------------------------------------------------------|---------------|------------|--------------|-----------|
| bert-base-uncased                                                     | 0.61          | 0.52       | 0.67         | 0.56      |
| stsb-roberta-large                                                    | 0.7           | 0.53       | 0.73         | 0.56      |
| all-distilroberta-v1                                                  | 0.65          | 0.52       | 0.65         | 0.55      |
| all-MiniLM-L6-v2                                                      | 0.64          | 0.52       | 0.69         | 0.54      |
| all-MiniLM-L12-v2                                                     | 0.66          | 0.53       | 0.67         | 0.56      |
| all-mpnet-base-v2                                                     | 0.66          | 0.51       | 0.68         | 0.52      |
| all-roberta-large-v1                                                  | 0.65          | 0.57       | 0.7          | 0.55      |
| paraphrase-MiniLM-L6-v2                                               | 0.66          | 0.52       | 0.73         | 0.54      |
| paraphrase-mpnet-base-v2                                              | 0.69          | 0.52       | 0.71         | 0.56      |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.65          | 0.53       | 0.7          | 0.55      |
| paraphrase-MiniLM-L12                                                 | 0.67          | 0.53       | 0.73         | 0.56      |
| paraphrase-MiniLM-L3                                                  | 0.65          | 0.53       | 0.69         | 0.56      |
| paraphrase-TinyBERT-L6                                                | 0.67          | 0.56       | 0.70         | 0.54      |
| paraphrase-distilroberta-base                                         | 0.66          | 0.56       | 0.72         | 0.54      |
| xlm-r-distilroberta                                                   | 0.68          | 0.52       | 0.72         | 0.566     |
| stsb-roberta-base-v2                                                  | 0.7           | 0.57       | 0.74         | 0.56      |
| stsb-roberta-large                                                    | 0.7           | 0.53       | 0.73         | 0.56      |
| stsb-distilroberta-base-v2                                            | 0.67          | 0.51       | 0.72         | 0.54      |
| stsb-mpnet-base-v2                                                    | 0.70          | 0.53       | 0.75         | 0.56      |
| multi-qa-distilbert-cos-v1                                            | 0.64          | 0.52       | 0.68         | 0.53      |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.65          | 0.53       | 0.70         | 0.55      |
| sentence-t5-base                                                      | 0.7           | 0.58       | 0.73         | 0.57      |
| sentence-t5-large                                                     | 0.71          | 0.59       | 0.73         | 0.59      |
| gtr-t5-base                                                           | 0.64          | 0.57       | 0.68         | 0.55      |
| gtr-t5-large                                                          | 0.67          | 0.59e      | 0.7          | 0.57      |
| nq-distilbert-base-v1                                                 | 0.64          | 0.52       | 0.68         | 0.54      |
| nli-distilroberta-base-v2                                             | 0.67          | 0.52       | 0.73         | 0.55      |
| sentence-transformers__paraphrase-multilingual-MiniLM-L12-v2          | 0.68          | 0.52       | 0.73         | 0.55      |
| sentence-transformers__paraphrase-multilingual-mpnet-base-v2          | 0.69          | 0.54       | 0.73         | 0.53      |
| sentence-transformers__distiluse-base-multilingual-cased-v2           | 0.63          | 0.55       | 0.66         | 0.53      |
| sentence-transformers__paraphrase-xlm-r-multilingual-v1               | 0.68          | 0.52       | 0.72         | 0.56      |
| sentence-transformers__distilbert-multilingual-nli-stsb-quora-ranking | 0.68          | 0.56       | 0.72         | 0.54      |
| sentence-transformers-quora-distilbert-multilingual                   | 0.68          | 0.56       | 0.72         | 0.54      |
| sentence-transformers__stsb-xlm-r-multilingual                        | 0.68          | 0.57       | 0.71         | 0.53      |
| sentence-transformers__distiluse-base-multilingual-cased              | 0.63          | 0.55       | 0.66         | 0.53      |

Table 1.11: Accuracy χρησιμοποιώντας τον αλγόριθμο ομαδοποίησης Kmeans

### 1.5.2 COMET

Όπως αναφέρθηκε προηγουμένως, το COMET εστιάζει στο task της μετάφρασης αλλά μπορεί να εφαρμοστεί και σε άλλα task του NLG. Έτσι, αρχίζουμε τα πειράματα για το task της μετάφρασης και έπειτα συνεχίζουμε στα υπόλοιπα tasks.

Για να πραγματοποιήσουμε μια λεπτομερή αξιολόγηση, διακρίναμε πρώτα τα σχετικά σημεία δεδομένων που αφορούν ειδικά τη μηχανική μετάφραση. Χρησιμοποιώντας αυτά τα σημεία δεδομένων, εφαρμόσαμε το μέτρο σύγκρισης COMET για να εξάγουμε μια ποσοτική βαθμολογία, η οποία κυμαίνεται από μηδέν έως ένα, με υψηλότερες τιμές να υποδηλώνουν ανώτερη ποιότητα μετάφρασης. Αυτή η μέθοδος αξιολόγησης παρείχε ένα αντικειμενικό μέτρο για το πόσο καλά τα μοντέλα απέδωσαν όσον αφορά την ακρίβεια, την ευχέρεια και τη συνολική ποιότητα της μετάφρασης. Τα αποτελέσματα αυτής της αξιολόγησης, όπως προέκυψαν από τα πειράματά μας, παρουσιάζονται στον Πίνακα 1.12, προσφέροντας πληροφορίες για τη συγκριτική απόδοση των υπό εξέταση μοντέλων.

| Σύνολο δεδομένων | Accuracy | Rho   |
|------------------|----------|-------|
| Val agnostic     | 0.82     | 0.718 |
| Val aware        | 0.71     | 0.64  |
| Test agnostic    | 0.78     | 0.6   |
| Test aware       | 0.75     | 0.61  |

Table 1.12: Accuracy και Συσχέτιση Spearman στο MT task χρησιμοποιώντας το COMET

Συνεχίζοντας τα πειράματά μας, εφαρμόζουμε την ίδια μεθοδολογία και στα υπόλοιπα tasks (DM and PG) παρουσιάζοντας τα αποτελέσματα στους Πίνακες 1.13 and 1.14

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.68     | 0.48 |
| Val aware     | 0.68     | 0.54 |
| Test agnostic | 0.69     | 0.5  |
| Test aware    | 0.65     | 0.46 |

Table 1.13: Accuracy και Συσχέτιση Spearman στο DM task χρησιμοποιώντας το COMET

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.68     | 0.54 |
| Val aware     | 0.8      | 0.27 |
| Test agnostic | 0.74     | 0.47 |
| Test aware    | 0.84     | 0.42 |

Table 1.14: Accuracy και Συσχέτιση Spearman στο PG task χρησιμοποιώντας το COMET

Αφού υπολογίσαμε τα αποτελέσματα για τα task ξεχωριστά, τα εννοποιήσαμε για να βγάλουμε τελικά αποτελέσματα αξιολόγησης για ολόκληρα τα σύνολα δεδομένων, τα οποία φαίνονται στον Πίνακα 1.15

### 1.5.3 Finetuning NLI μοντέλο

Σε αυτή την ενότητα, περιγράφουμε τα πειράματα που πραγματοποιήθηκαν για την αξιολόγηση της αποτελεσματικότητας της περαιτέρω εκπαίδευσης μοντέλων συμπερασμού φυσικής γλώσσας (NLI) στην ανίχνευση hallucination που παράγονται από μεγάλα γλωσσικά μοντέλα (LLM). Χρησιμοποιήσαμε προ-εκπαιδευμένα μοντέλα NLI και τα εκπαιδεύσαμε περαιτέρω χρησιμοποιώντας το σύνολο δεδομένων validation του SHROOM. Ευθυγραμμίζοντας αυτά τα μοντέλα με το έργο της ανίχνευσης hallucination, τα εκπαιδεύσαμε να αξιολογούν τη

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.74     | 0.55 |
| Val aware     | 0.58     | 0.17 |
| Test agnostic | 0.73     | 0.53 |
| Test aware    | 0.57     | 0.15 |

Table 1.15: Accuracy και Συσχέτιση Spearman για όλα τα task χρησιμοποιώντας το COMET

σχέση μεταξύ μιας πρότασης-στόχου (βασική αλήθεια) και μιας υπόθεσης (έξοδος από το LLM). Στη συνέχεια, η υπόθεση ταξινομείται είτε ως συνεπαγόμενη, είτε ως αντιφατική, είτε ως ουδέτερη όσον αφορά τον στόχο, παρέχοντας ένα πλαίσιο για την ανίχνευση hallucination.

Χρησιμοποιήσαμε ένα προ-εκπαιδευμένο μοντέλο NLI από τη βιβλιοθήκη Hugging Face, συγκεκριμένα αρχιτεκτονικής mDeBERTa-v3-base. Αυτό το μοντέλο, μέρος της τρίτης επανάληψης της οικογένειας DeBERTa (Decoding-enhanced BERT with disentangled attention), εκπαιδεύτηκε σε ένα μεγάλο πολύγλωσσο σύνολο δεδομένων, καθιστώντας το άριστο υποψήφιο για την εργασία μας. Η προ-εκπαίδευση σε ένα τόσο μεγάλης κλίμακας σύνολο δεδομένων επέτρεψε στο μοντέλο να συλλάβει ένα ευρύ φάσμα γλωσσικών χαρακτηριστικών και μοτίβων σε διάφορες γλώσσες, προσφέροντας μια σταθερή βάση για την επεξεργασία ενός ευρέος φάσματος γλωσσικών δομών. Αυτό ήταν ζωτικής σημασίας για τον χειρισμό της πολυπλοκότητας και της μεταβλητότητας των δεδομένων με τα οποία εργαστήκαμε, τόσο όσον αφορά τη γλωσσική ποικιλομορφία όσο και τα χαρακτηριστικά περιεχομένου.

Το μοντέλο NLI είχε ήδη εκπαιδευτεί στο πολύγλωσσο σύνολο δεδομένων multilingual-nli-26lang-2mil7 και στο σύνολο δεδομένων επικύρωσης XNLI, τα οποία περιείχαν τρεις ετικέτες: «συνεπαγωγή», «ουδέτερη» και «αντίφαση». Κατά τη διάρκεια της εκπαίδευσης, αντιστοιχίσαμε την ετικέτα «Παραισθήσεις» στην «αντίφαση» και την ετικέτα «Όχι παραισθήσεις» στην «συνεπαγωγή», δημιουργώντας μια δυαδική ταξινόμηση για να αναπαραστήσουμε με σαφήνεια την παραισθησιογόνο φύση του περιεχομένου. Αυτός ο μετασχηματισμός απλοποίησε τη διαδικασία εκμάθησης παρέχοντας διακριτές ετικέτες στο μοντέλο για τη διάκριση μεταξύ παραισθησιογόνων και μη παραισθησιογόνων περιπτώσεων.

Τελικά, ο συνδυασμός ενός ισχυρού προ-εκπαιδευμένου μοντέλου με την εξειδικευμένη για κάθε εργασία τελειοποίηση μας επέτρεψε να αναπτύξουμε ένα εξαιρετικά εξειδικευμένο εργαλείο για την ανίχνευση hallucination σε παραγόμενες απαντήσεις. Η ικανότητα του μοντέλου να λειτουργεί αποτελεσματικά στο πλαίσιο της εργασίας μας ενισχύθηκε, διασφαλίζοντας ότι θα μπορούσε να κάνει τις ακριβείς διακρίσεις που απαιτούνται για τη βελτίωση της συνολικής ποιότητας και αξιοπιστίας του παραγόμενου περιεχομένου.

Μετά την εκπαίδευση, οι προβλέψεις του μοντέλου αξιολογήθηκαν χρησιμοποιώντας μια βαθμολογία συνεπαγωγής και εφαρμόστηκε ένα προσεκτικά επιλεγμένο κατώφλι για τη διάκριση μεταξύ hallucination και μη hallucination. Μέσω πειραματισμού, διαπιστώσαμε ότι ένα κατώφλι 0,8 βελτιστοποίησε την ακρίβεια του δοκιμαστικού μας συνόλου. Επιπλέον, για τον υπολογισμό του ποσοστού hallucination για κάθε σημείο δεδομένων, αφαιρέσαμε το ποσοστό συνεπαγωγής από το 1.

Οι υπερπαραμέτροι που χρησιμοποιήθηκαν για το finetuning του μοντέλου NLI ταυτίζονται με εκείνες που χρησιμοποιήθηκαν κατά την εκπαίδευση του αρχικού μοντέλου. Η επιλογή των υπερπαραμέτρων ακολούθησε μια σειρά πειραμάτων, τα οποία απέδωσαν σημαντικά χαμηλότερα επίπεδα ακρίβειας. Ορισμένα από τα πειράματα εμφανίζονται στον Πίνακα 1.16, ενώ μια λεπτομερής περιγραφή των παραμέτρων που χρησιμοποιήθηκαν για την εκπαίδευση αυτού του μοντέλου NLI περιγράφεται στον Πίνακα 1.17.

Τα αποτελέσματα πριν και μετά το finetuning φαινεται στον Πίνακα 1.18

### Fine-tune προ-εκπαιδευμένου μοντέλο για ανίχνευση hallucination

Το finetuned μοντέλο υποβλήθηκε σε μια ολοκληρωμένη διαδικασία εκπαίδευσης και αξιολόγησης, αξιοποιώντας τα δεδομένα SHROOM που παρείχαν οι διοργανωτές του task. Τα δεδομένα αυτά ήταν απαραίτητα για την εκπαίδευση του μοντέλου ώστε να ανιχνεύει με ακρίβεια τα hallucinations στις παραγόμενες εξόδους. Συγκεκριμένα, το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας το validation σύνολο δεδομένων, με τις ετικέτες για την

| epochs | lr    | warmup ratio | weight decay | accuracy |
|--------|-------|--------------|--------------|----------|
| 5      | 2e-05 | 0.06         | 0.01         | 0.83     |
| 10     | 2e-06 | 0.1          | 0.01         | 0.75     |
| 5      | 2e-04 | 0.01         | 0.05         | 0.53     |
| 5      | 2e-05 | 0.05         | 0.001        | 0.8      |
| 5      | 2e-06 | 0.08         | 0.1          | 0.79     |

Table 1.16: Accuracy στο trial-set για τον καθορισμό των υπερπαραμέτρων. Η πρώτη σειρά δείχνει τις υπερπαραμέτρους που επιλέχθηκαν.

| Hyperparameter | Value                          |
|----------------|--------------------------------|
| train dataset  | validation set (1,000 samples) |
| learning rate  | 2e-05                          |
| epochs         | 5                              |
| warm-up ratio  | 0.06                           |
| weight decay   | 0.01                           |

Table 1.17: Υπερπαραμέτροι που χρησιμοποιήθηκαν για το fine-tuning

|                   | test agn | test aware | val agn | val aware | trial |
|-------------------|----------|------------|---------|-----------|-------|
| Before finetuning | 0.744    | 0.736      | 0.772   | 0.735     | 0.812 |
| After fine-tuning | 0.751    | 0.77       | -       | -         | 0.83  |

Table 1.18: Accuracy των nli μοντέλων πριν και μετά το finetuning

ανίχνευση hallucination, και η απόδοσή του αξιολογήθηκε στη συνέχεια χρησιμοποιώντας ένα δοκιμαστικό σύνολο για να διασφαλιστεί ότι γενικεύεται καλά σε καινούρια δεδομένα.

Στη φάση της προεπεξεργασίας, κάθε σημείο δεδομένων προετοιμάστηκε προσεκτικά για την εισαγωγή στο μοντέλο. Εξαγάγαμε δύο βασικά στοιχεία από κάθε περίπτωση: την υπόθεση (*hyp*) και τον στόχο (*tgt*). Το *hyp* αντιπροσώπευε την παραγόμενη έξοδο από το μοντέλο που έπρεπε να αξιολογηθεί για hallucinations, ενώ το *tgt* χρησίμευε ως αναφορά ή βασική αλήθεια, παρέχοντας τη βάση για σύγκριση. Αυτά τα στοιχεία τροφοδοτήθηκαν στη συνέχεια στο μοντέλο για να του επιτρέψουν να αξιολογήσει εάν η παραγόμενη έξοδος ευθυγραμμίζεται με το κείμενο αναφοράς ή αποκλίνει με τρόπους που υποδηλώνουν hallucinations.

Για να ενισχύσουμε περαιτέρω την απόδοση του μοντέλου όσον αφορά τόσο την ακρίβεια όσο και την ικανότητά του να προβλέπει την πιθανότητα hallucination ( $p(\text{'Hallucination'})$ ), εφαρμόσαμε μια στρατηγική διπλής εκπαίδευσης. Αυτή η στρατηγική ήταν καθοριστικής σημασίας για τη βελτιστοποίηση των προβλέψεων του μοντέλου και τη βελτίωση της ευελιξίας του.

Στην πρώτη επανάληψη, εκπαιδεύσαμε το μοντέλο χρησιμοποιώντας δυαδικές ετικέτες: '0' για hallucinations και '1' για μη hallucinations. Αυτή η δυαδική ταξινόμηση επέτρεψε στο μοντέλο να λαμβάνει απλές αποφάσεις σχετικά με το αν μια δεδομένη υπόθεση περιείχε ή όχι παραισθησιογόνο περιεχόμενο.

Στη δεύτερη επανάληψη, ακολουθήσαμε μια πιο διαφοροποιημένη προσέγγιση, χρησιμοποιώντας ετικέτες float που αντιπροσώπευαν το συμπλήρωμα της πιθανότητας hallucination, ή  $(1-p(\text{'Hallucination'}))$ . Αυτό επέτρεψε στο μοντέλο να παράγει μια πιο λεπτομερή βαθμολογία πιθανότητας, αντικατοπτρίζοντας την πιθανότητα hallucination στην παραγόμενη έξοδο και όχι απλώς αποδίδοντας μια δυαδική ετικέτα. Χρησιμοποιώντας αυτές τις συνεχείς τιμές, το μοντέλο μπορούσε να αποτυπώσει καλύτερα το βαθμό αβεβαιότητας ή εμπιστοσύνης που σχετίζεται με τις προβλέψεις του.

Αυτή η προσέγγιση διπλής εκπαίδευσης προσέφερε αρκετά πλεονεκτήματα. Πρώτον, επέτρεψε στο μοντέλο να παρέχει ταυτόχρονα μια δυαδική ετικέτα που υποδεικνύει την παρουσία ή την απουσία hallucination και μια βαθμολογία πιθανότητας που ποσοτικοποιεί την πιθανότητα hallucination. Αυτό μας επέτρεψε να εξισορροπήσουμε μεταξύ της λήψης κατηγορικών αποφάσεων σχετικά με τα hallucinations και ταυτόχρονα να αποκτήσουμε μια βαθύτερη εικόνα της εμπιστοσύνης του μοντέλου στις προβλέψεις του. Η βαθμολογία πιθανότητας ήταν ιδιαίτερα χρήσιμη για σενάρια όπου χρειαζόταν μια πιο λεπτομερής κατανόηση των αποτελεσμάτων του μοντέλου, όπως

όταν απαιτούνταν περαιτέρω ανθρώπινη αναθεώρηση ή βελτίωση.

Συνολικά, αυτός ο συνδυασμός δυαδικών και πιθανοτικών αποτελεσμάτων, που επιτεύχθηκε μέσω της διαδικασίας διπλής εκπαίδευσης, ενίσχυσε την προσαρμοστικότητα και την αποτελεσματικότητα του μοντέλου στην ανίχνευση hallucination. Επίσης, παρείχε μια πιο ολοκληρωμένη κατανόηση των αποτελεσμάτων που παράγονται από το μοντέλο, διασφαλίζοντας ότι η ανίχνευση hallucination ήταν τόσο ακριβής όσο και διορατική. Οι υπερπαράμετροι για τη λεπτομερή ρύθμιση περιγράφονται αναλυτικά στον Πίνακα 1.19.

Τα αποτελέσματα φαίνονται στο Πίνακα 1.20

| Hyperparameter   | Value                          |
|------------------|--------------------------------|
| train dataloader | validation set (1,000 samples) |
| evaluator        | trial set (80 samples)         |
| epochs           | 5                              |
| evaluation steps | 10,000                         |
| warm-up steps    | 10% of train data for warm-up  |

Table 1.19: Υπερπαράμετροι που χρησιμοποιήθηκαν για το finetuning του μοντέλου

| acc<br>test-agn | rho<br>test-agn | acc<br>test-aware | acc<br>test-aware | trial |
|-----------------|-----------------|-------------------|-------------------|-------|
| 0.778           | 0.668           | 0.795             | 0.685             | 0.85  |

Table 1.20: Αποτελέσματα από το Fine-tune hal-detect model's

#### 1.5.4 Ταξινομητής Ψηφοφορίας

Στην τελική μας προσέγγιση, υιοθετήσαμε μια τεχνική συνόλου γνωστή ως Voting Classifier για να βελτιώσουμε την ευρωστία και την ακρίβεια του προγνωστικού μας μοντέλου. Ο ταξινομητής ψηφοφορίας λειτουργεί με βάση την αρχή της συγγέντρωσης των αποτελεσμάτων πολλαπλών μεμονωμένων ταξινομητών για να καταλήξει σε μια τελική απόφαση. Κάθε ταξινομητής εντός του συνόλου παράγει την πρόβλεψή του για την κλάση εξόδου με βάση τα δεδομένα εισόδου που του δίνονται και ο ταξινομητής ψηφοφορίας ενοποιεί αυτές τις προβλέψεις. Η τελική προβλεπόμενη κλάση καθορίζεται με ψηφοφορία πλειοψηφίας, με την οποία η κλάση που λαμβάνει τον μεγαλύτερο αριθμό ψήφων από τους επιμέρους ταξινομητές επιλέγεται ως τελική έξοδος.

Η δύναμη αυτής της προσέγγισης συνόλου έγκειται στην ικανότητά της να αξιοποιεί τα μοναδικά πλεονεκτήματα κάθε συστατικού ταξινομητή. Στην περίπτωση μας, ενσωματώσαμε διάφορα μοντέλα, από προηγούμενα κεφάλαια. Συγκεκριμένα, χρησιμοποιήσαμε το προ-εκπαιδευμένο μοντέλο ανίχνευσης hallucination, τη fine-tuned εκδοχή του και το fine-tuned μοντέλο NLI, καθένα από τα οποία εκπαιδεύτηκε για να αντιμετωπίσει το έργο από διαφορετικές οπτικές γωνίες και μέσω διαφορετικών μεθοδολογιών. Συνδυάζοντας αυτά τα διαφορετικά μοντέλα, μπορέσαμε να αξιοποιήσουμε τα επιμέρους πλεονεκτήματα κάθε προσέγγισης. Ορισμένοι ταξινομητές μπορεί να είναι πιο κατάλληλοι για συγκεκριμένες πτυχές ή χαρακτηριστικά των δεδομένων- έτσι, οι προβλέψεις τους αντικατοπτρίζουν αυτές τις αποχρώσεις. Ωστόσο, επιτρέποντας σε πολλαπλά μοντέλα να συνεισφέρουν στη διαδικασία λήψης αποφάσεων, το σύνολο βοηθά στην εξισορρόπηση πιθανών αδυναμιών ή προκαταλήψεων που μπορεί να προκύψουν όταν στηριζόμαστε σε ένα μόνο μοντέλο.

Τα αποτελέσματα αυτής της μεθόδου παρουσιάζονται στον Πίνακα 1.21, ο οποίος πέτυχε την υψηλότερη ακρίβεια από όλες τις υποβολές μας για τον διαγωνισμό Semeval.

#### 1.5.5 Prompting

Όπως συζητήθηκε προηγουμένως, η τελική μας προσέγγιση επικεντρώθηκε στην εφαρμογή του Prompt Engineering. Για την έναρξη αυτής της μεθόδου, σχεδιάσαμε και εφαρμόσαμε zero-shot prompting, αναπτύσσοντας την σε μια σειρά μοντέλων, συμπεριλαμβανομένων των LLaMA-2 (7B, 13B και 70B), LLaMA-3 70B και GEMMA, μεταξύ άλλων. Στη συνέχεια, επεκτείναμε την προσέγγισή μας ώστε να συμπεριλάβουμε και few-shot prompting και διερευνήσαμε περαιτέρω τη συλλογιστική της αλυσίδας σκέψης (Chain of Thought-CoT)

| Method                                | acc.↑        | rho↑         |
|---------------------------------------|--------------|--------------|
| <b>Model-aware</b>                    |              |              |
| Baseline Model                        | 0.745        | 0.488        |
| Fine-tune hal-detect model            | 0.795        | 0.685        |
| NLI model                             | 0.77         | 0.591        |
| Voting Classifier-majority vote       | <b>0.799</b> | 0.691        |
| Voting Classifier-averaged percentage | <b>0.799</b> | <b>0.693</b> |
| <b>Model-agnostic</b>                 |              |              |
| Baseline Model                        | 0.697        | 0.402        |
| Fine-tune hal-detect model            | 0.778        | 0.668        |
| NLI model                             | 0.751        | 0.548        |
| Voting Classifier-majority vote       | <b>0.78</b>  | 0.632        |
| Voting Classifier-averaged percentage | <b>0.78</b>  | <b>0.643</b> |

Table 1.21: Αποτελέσματα για την μέθοδο του Ταξινομητή Ψηφοφορίας. Τα έντονα γράμματα δηλώνουν τα καλύτερα αποτελέσματα

ενσωματώνοντας τ prompts που απεικονίζεται στον πίνακα 1.25. Τα αποτελέσματα κάθε αντίστοιχης μεθόδου -zero-shot, few-shot και CoT- παρουσιάζονται αντίστοιχα στον Πίνακα 1.22, στον Πίνακα 1.23 και στον Πίνακα 1.24, αντίστοιχα. Τα αποτελέσματα αυτά προσφέρουν μια συγκριτική ανάλυση των επιδόσεων σε διαφορετικά μοντέλα και τεχνικές prompting.

| Model       | Acc test-agnostic | Acc test-aware |
|-------------|-------------------|----------------|
| Llama-2 7b  | 0.60              | 0.62           |
| Llama-2 13b | 0.64              | 0.67           |
| Llama-2 70b | <b>0.77</b>       | <b>0.75</b>    |
| Llama-3 70b | <b>0.77</b>       | <b>0.75</b>    |
| Gemma 2b    | 0.60              | 0.61           |
| Gemma 7b    | 0.74              | 0.7            |
| Mistral 7b  | 0.46              | 0.47           |

Table 1.22: Αποτελέσματα για Zero-shot prompting

| Model       | Acc test-agnostic | Acc test-aware |
|-------------|-------------------|----------------|
| Llama-2 7b  | **                | **             |
| Llama-2 13b | 0.42              | 0.51           |
| Llama-2 70b | 0.62              | 0.64           |
| Llama-3 8b  | **                | **             |
| Llama-3 70b | 0.62              | 0.65           |
| Gemma 2b    | 0.57              | 0.55           |
| Gemma 7b    | **                | **             |
| Mistral 7b  | 0.5               | 0.54           |

Table 1.23: Αποτελέσματα για Few-shot prompting

**Παρατηρήσεις:** Τα αποτελέσματα από τα πειράματά μας δείχνουν ότι το πιο αποτελεσματικό μοντέλο σε αυτό το πλαίσιο είναι το LLaMA-2 70B, ιδιαίτερα όταν χρησιμοποιείται το zero-shot. Αυτή η ανώτερη απόδοση μπορεί να αποδοθεί στην προηγμένη αρχιτεκτονική του μοντέλου, σε συνδυασμό με την απλοποίηση του prompting, που επιτρέπει την ακριβέστερη επεξεργασία της εισόδου. Επιπλέον, παρατηρήσαμε μια τάση σε

| Model       | Acc test-agnostic | Acc test-aware |
|-------------|-------------------|----------------|
| Llama-2 7b  | 0.63              | 0.65           |
| Llama-2 13b | 0.58              | 0.56           |
| Llama-2 70b | 0.57              | 0.61           |
| Llama-3 8b  | **                | **             |
| Llama-3 70b | 0.58              | 0.62           |
| Gemma 2b    | 0.59              | 0.63           |
| Gemma 7b    | 0.60              | 0.63           |
| Mistral 7b  | 0.49              | 0.46           |

Table 1.24: Αποτελέσματα για CoT prompting

| Method              | Prompt                                                                                                                                                                                                                                                                                                                                                              |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Zero-shot Prompting | [Sentence 1]: Freedom is worthless.<br>[Sentence 2]: Freedom is invaluable.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.                                                                                                                                                                                                             |
| Few-shot            | In the following example, the Sentence 1 is supported by the Sentence 2:<br>example_1<br>In the following example, the Sentence 1 is not supported by the Sentence 2:<br>example_2<br>Now, it is your turn to evaluate the following example.<br>[Sentence 1]: hyp<br>[Sentence 2]: tgt<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No. |
| Chain-of-Thought    | [Sentence 1]: Freedom is worthless.<br>[Sentence 2]: Freedom is invaluable.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>Explain your answer.                                                                                                                                                                                     |

Table 1.25: Prompts που χρησιμοποιήθηκαν για την κάθε μέθοδο

μικρότερα μοντέλα, όπως το LLaMA-7B, να παρουσιάζουν μεροληπτική συμπεριφορά, αποδίδοντας σταθερά μια ενιαία ετικέτα σε όλα τα σημεία δεδομένων. Συγκεκριμένα, το μοντέλο ταξινομούσε κάθε περίπτωση είτε ως «Hallucination» είτε ως «Not Hallucination», ανάλογα με τη διαμόρφωση της εκπαίδευσής του, οδηγώντας σε διογκωμένα ποσοστά recall, όπως φαίνεται στον Πίνακα 1.27. Επιπλέον, τα ευρήματά μας υποδεικνύουν ότι καθώς αυξανόταν η πολυπλοκότητα του prompt, η κατανόηση και η συνολική ακρίβεια του μοντέλου μειώνονταν σημαντικά. Αυτό αναδεικνύει την ευαισθησία αυτών των μοντέλων στην πολυπλοκότητα των prompt, με τα υπερβολικά περίπλοκα prompting να οδηγούν σε μειωμένη απόδοση και μειωμένη κατανόηση της εκάστοτε εργασίας. Επιπλέον, στον Πίνακα 1.23 και στον Πίνακα 1.24, οι καταχωρήσεις που σημειώνονται με \*\* υποδεικνύουν

περιπτώσεις όπου οι απαντήσεις του μοντέλου ήταν εντελώς άσχετες με την είσοδο, καθιστώντας αδύνατη την εξαγωγή έγκυρης απάντησης. Μερικά παραδείγματα τέτοιων απαντήσεων παρουσιάζονται στον Πίνακα 1.26

| Μοντέλο    | Είσοδος                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Απάντηση                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Llama-3 8b | "[Sentence 1]: Tom thought Mary probably wouldn't give up French.<br>[Sentence 2]: Tom thought Mary would probably fail her French test.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>Explain your answer.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <a href="https://www.duolingo.com/bookmark/1206a4">https://www.duolingo.com/bookmark/1206a4</a>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| gemma 7b   | First, you will be provided with 2 examples with the correct answer, similar to the question you will need to answer.<br>[Sentence 1]: Freedom is worthless.<br>[Sentence 2]: Freedom is invaluable.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>[Correct Answer]: No, Sentence 1 is not supported by the Sentence 2.<br>[Sentence 1]: This isn't your house anymore.<br>[Sentence 2]: You don't live here anymore.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>[Correct Answer]: Yes, Sentence 1 is supported by the Sentence 2.<br>Now, it is your turn to evaluate the following example<br>[Sentence 1]: Can anyone corroborate that?<br>[Sentence 2]: Can anyone back you up on that? Sentence 1 is not supported by the Sentence 2 Answer with Yes or No. | ## The answers for each sentence:<br>**Freedom and Can Anyone Corroborate/Back Me Up? **<br>### Sentences about freedom : (sentence one) - FREEDOM IS WORTHLESS / (sentence two)-FREEDOM HAS VALUE . The sentences are saying different things despite using seemingly synonymous phrases like 'freedom', valuable & useless'. They differ in tone of voice , purpose as well context which significantly impacts their meanings even though they both deal broadly around concepts related directly opposite from valuedness versus worthlessness.. Therefore I believe these statements do NOT carry identical semantic weight hence my response was NO. **_***sentences regarding living space:** - [Sentence One] THIS AIN'T YOUR HOUSE ANYMORE./-[Sentence Two]-YOU DON'T LIVE HERE AGAIN AND THE ANSWER FOR BOTH SENTENCES ARE YES BECAUSE THEY CARRY SAME MEANING BUT JUST WITH DIFFERENT WAYS OF SAYINGS THAT STILL HAVE A LIKE TO ONE Another. |

Table 1.26: Απαντήσεις άσχετες με την είσοδο του μοντέλου

Στην τελική μας προσπάθεια να βελτιώσουμε τις επιδόσεις, επικεντρωθήκαμε στην τελειοποίηση των zero-shot που χρησιμοποιήθηκαν στο LLaMA-2 70B, που είχε προηγουμένως επιδείξει την υψηλότερη ακρίβεια από όλα τα μοντέλα. Για κάθε συγκεκριμένο task, προσαρμόσαμε τα prompts ώστε να ευθυγραμμιστούν καλύτερα με τις απαιτήσεις της εργασίας. Τα τροποποιημένα prompts για κάθε task περιγράφονται λεπτομερώς στον πίνακα 1.28. Τα αποτελέσματα μετά από αυτές τις προσαρμογές παρέχονται στον πίνακα 1.29, το οποίο αποδεικνύει για άλλη μια φορά ότι τα υψηλότερα αποτελέσματα προέρχονται από τα απλά prompts.

| Model       | Recall test-agnostic<br>Hallucination/Not<br>Hallucination | Recall test-aware<br>Hallucination/Not<br>Hallucination |
|-------------|------------------------------------------------------------|---------------------------------------------------------|
| Llama-2 7b  | 0.25/0.85                                                  | 0.39/0.76                                               |
| Llama-2 13b | 0.29/0.82                                                  | 0.29/0.82                                               |
| Llama-2 70b | 0.5/0.95                                                   | 0.52/0.88                                               |
| Llama-3 70b | 0.53/0.93                                                  | 0.41/0.95                                               |
| Gemma 2b    | 0.04/0.99                                                  | 0.03/0.96                                               |
| Gemma 7b    | 0.49/0.91                                                  | 0.42/0.87                                               |
| Mistral 7b  | 0.5/0.93                                                   | 0.4/0.8                                                 |

Table 1.27: Recall for each dataset of zero-shot prompt

| Task | Prompt                                                                                                                                                                                                                                                                         |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DM   | [Sentence 1]: src<br>[Sentence 2]: hyp<br>[Sentence 3]: tgt<br>A machine was given the task of giving the definition of Sentence 1. The machine's answer was Sentence 2 while the expected answer was Sentence 3. Did the machine answer correctly?<br>Answer with Yes or No.  |
| PG   | [Sentence 1]: src<br>[Sentence 2]: hyp<br>A machine was given the task of giving a paraphrase of Sentence 1. The machine's answer was Sentence 2. Did the machine answer correctly?<br>Answer with Yes or No.                                                                  |
| MT   | [Sentence 1]: src<br>[Sentence 2]: hyp<br>[Sentence 3]: tgt<br>A machine was given the task of giving the translation of Sentence 1. The machine's answer was Sentence 2 while the expected answer was Sentence 3. Did the machine answer correctly?<br>Answer with Yes or No. |

Table 1.28: Prompts που χρησιμοποιήθηκαν για κάθε Task

| Model       | Accuracy test-agnostic | Accuracy test-aware |
|-------------|------------------------|---------------------|
| Llama-3 70b | 0.66                   | 0.63                |

Table 1.29: Αποτελέσματα για τα διαφορετικά prompts του κάθε task

## 1.6 Σύνοψη

Η παρούσα διπλωματική ασχολήθηκε με το κρίσιμο ζήτημα των hallucination στα Μεγάλα Γλωσσικά Μοντέλα (LLMs), ιδίως στο πλαίσιο των εργασιών παραγωγής φυσικής γλώσσας (NLG), όπου τέτοια σφάλματα μπορούν να υπονομεύσουν την αξιοπιστία των συστημάτων NLG. Η ευρεία υιοθέτηση των LLMs σε διάφορους τομείς έχει αυξήσει τη σημασία της ανάπτυξης ισχυρών μεθόδων για την ανίχνευση και τον μετριασμό των hallucination, ιδίως σε περιβάλλοντα black-box όπου η εσωτερική λειτουργία των μοντέλων παραμένει άγνωστη. Μέσω της συμμετοχής στο SemEval-2024 Task 6 (SHROOM), η έρευνα αυτή διερεύνησε νέες προσεγγίσεις για την ανίχνευση hallucination που εφαρμόζονται σε απαιτητικούς τομείς όπως η μηχανική μετάφραση, η μοντελοποίηση ορισμών και η παραγωγή παραφράσεων.

Οι βασικές συνεισφορές αυτής της εργασίας περιλαμβάνουν τη λεπτομερή ρύθμιση προ-εκπαιδευμένων μοντέλων σε σύνολα δεδομένων ανίχνευσης hallucination και εξαγωγής συμπερασμάτων φυσικής γλώσσας, επιτυγχάνοντας ανώτερες επιδόσεις σε σύγκριση με τα βασικά μοντέλα. Η ενσωμάτωση αυτών των μοντέλων σε ένα πλαίσιο ταξινομητή ψηφοφορίας έχει επιδείξει ανταγωνιστική ακρίβεια στην ανίχνευση hallucination σε σύνολα δεδομένων που είτε γνωρίζουμε το μοντέλο είτε όχι. Επιπλέον, η παρούσα έρευνα εισήγαγε μια νέα προσέγγιση prompting για τη βελτίωση της ερμηνευσιμότητας, καθοδηγώντας τα μοντέλα να εντοπίζουν σημασιολογικές ασυνέπειες στα παραγόμενα αποτελέσματα. Χρησιμοποιώντας μετρικές ομοιότητας προτάσεων, η μεθοδολογία επέτρεψε περαιτέρω μια ποσοτική αξιολόγηση των hallucination, βασισμένη στη σύγκριση μεταξύ των παραγόμενων κειμένων και των κειμένων εισόδου.

Τα αποτελέσματα που επιτεύχθηκαν στην παρούσα διατριβή υπογραμμίζουν τη δυνατότητα συνδυασμού πολλαπλών στρατηγικών, συμπεριλαμβανομένων του finetuning, των μεθόδων συνόλου και των σημασιολογικών μετρικών, για τη βελτίωση της ακρίβειας ανίχνευσης hallucination. Η ανάλυση των μοτίβων σφαλμάτων σε διαφορετικές εργασίες παρείχε επίσης πολύτιμες πληροφορίες για τη φύση των hallucination, αποκαλύπτοντας τους διαφορετικούς βαθμούς πρόκλησης που θέτουν οι διαφορετικοί τύποι εργασιών NLG. Για παράδειγμα, η ανώτερη απόδοση που παρατηρήθηκε στη Δημιουργία Παραφράσεων συνάδει με τις εγγενείς ιδιότητες της εργασίας, ενώ η χαμηλότερη ακρίβεια στη Μοντελοποίηση ορισμών αντανακλά τη μεγαλύτερη δυσκολία της ακριβούς μοντελοποίησης ορισμών. Επιπλέον, η μελέτη των σφαλμάτων με βάση την πιθανότητα ύπαρξης hallucination ανέδειξε την δυσκολία που υπάρχει στον χαρακτηρισμό των hallucination ακόμα και από ανθρώπους.

Εν κατακλείδι, η παρούσα έρευνα προωθεί την τεχνολογία ανίχνευσης hallucination για LLMs προτείνοντας μια σειρά μεθόδων που είναι τόσο αποτελεσματικές όσο και υπολογιστικά αποδοτικές μέσα σε μια διάταξη black box. Τα ευρήματα θέτουν τα θεμέλια για μελλοντικές εργασίες που αποσκοπούν στην ενίσχυση της αξιοπιστίας των συστημάτων που βασίζονται σε LLM με την αντιμετώπιση της εγγενούς πρόκλησης των hallucination.

### Μελλοντικές εργασίες

1. **Τεχνικές προσαρμοσμένες στις εργασίες:** Ανάπτυξη εξειδικευμένων μεθόδων προσαρμοσμένων στα χαρακτηριστικά κάθε εργασίας παραγωγής φυσικής γλώσσας (NLG). Λαμβάνοντας υπόψη τη μοναδική φύση και τις απαιτήσεις εργασιών όπως η περίληψη, η μετάφραση και η παραγωγή διαλόγων, είναι δυνατόν να δημιουργηθούν πιο αποτελεσματικές προσεγγίσεις ανίχνευσης hallucination.
2. **Πολυτροπικές προσεγγίσεις:** Επέκταση του πεδίου της έρευνας ώστε να συμπεριλάβει και άλλους τομείς, όπως η απάντηση ερωτήσεων με βάση κάποιες εικόνες και η παραγωγή λεζάντες εικόνων. Η διερεύνηση της ανίχνευσης hallucination σε πολυτροπικές εργασίες, όπου το κείμενο παράγεται σε συνδυασμό με άλλες μορφές δεδομένων, θα μπορούσε να οδηγήσει σε πιο ολοκληρωμένες και ισχυρές μεθόδους.
3. **Ενσωμάτωση εξωτερικών βάσεων γνώσης:** Διερεύνηση της χρήσης του Retrieval-Augmented Generation (RAG) και άλλων τεχνικών για τη βελτίωση της επαλήθευσης της ακρίβειας των γεγονότων στο παραγόμενο περιεχόμενο. Η αξιοποίηση εξωτερικών βάσεων γνώσης, όπως τα Wikidata ή βάσεις δεδομένων ειδικών τομέων, θα μπορούσε να βοηθήσει στην επικύρωση της αληθοφάνειας των παραγόμενων αποτελεσμάτων.
4. **Εξηγήσιμες τεχνικές τεχνητής νοημοσύνης (XAI):** Ανάπτυξη μεθόδων επεξηγηματικότητας για την απόκτηση βαθύτερης κατανόησης της φύσης των hallucination και την κατανόηση των υποκείμενων λόγων για την εμφάνισή τους.
5. **Αυτόματη διάθροιση hallucination:** Πέρα από την ανίχνευση hallucination, διερεύνηση μεθόδων

για την αυτόματη διόρθωση ή αναθεώρηση των αποτελεσμάτων των μοντέλων ώστε να είναι συνεπή ως προς τα γεγονότα.

# Chapter 2

## Introduction

In recent years, the field of Natural Language Processing (NLP) has been significantly transformed by the advent and widespread adoption of Large Language Models (LLMs). These models have not only revolutionized the way natural language tasks are approached but have also raised new concerns regarding their reliability and the accuracy of the outputs they generate. One of the most pressing issues confronting the NLP community is the phenomenon of hallucinations in LLMs, wherein models produce information that is not grounded in the input data or factual reality. This issue has garnered considerable attention in the research community, as it directly affects the trustworthiness and utility of LLM-based systems across a wide array of applications [44],[20], [19], [52], [53]. Beyond hallucinations, LLMs have also been studied in areas such as reasoning, where researchers investigate the models’ capabilities in interpreting complex puzzles and riddles, further exploring the depth of LLMs’ understanding [40], [39], [15]. Despite the impressive capabilities of LLMs, these hallucinations present a significant challenge, as they undermine the dependability of systems designed to assist in decision-making, content generation, and more. Recent studies have suggested that hallucinations are not merely incidental or avoidable errors but are instead intrinsic to the nature of LLMs, given their probabilistic underpinnings. These findings indicate that hallucinations may be inevitable to some extent, regardless of the sophistication of the underlying model architecture [51]. As such, addressing the issue of hallucinations, particularly in the context of Natural Language Generation (NLG) tasks, has become a critical area of focus. Furthermore, the field of Explainable AI (XAI) has contributed valuable insights for understanding model behavior, especially through counterfactual explanations that reveal how minor input changes can affect outputs. This research on counterfactuals has been pivotal in exploring ways to control and predict LLM outputs, providing frameworks that enhance transparency and reliability in NLP systems [31],[14]. However, despite the growing body of research on the topic, there remains a notable scarcity of datasets and benchmarks specifically designed to evaluate and mitigate hallucinations in NLG tasks [26, 6, 7, 38].

In response to this need, the SemEval-2024 Task 6 (SHROOM: Shared-task on Hallucinations and Related Observable Overgeneration Mistakes) [35] has been introduced to address the challenge of identifying semantically unrelated generations in NLP tasks. This task encompasses several challenging domains, including Machine Translation, Definition Modelling, and Paraphrase Generation. What makes SHROOM particularly demanding is the inclusion of both known and unknown model settings, pushing the boundaries of current hallucination detection strategies and requiring the development of approaches that are robust across varying levels of model transparency.

This thesis aims to explore effective and broadly applicable strategies for hallucination detection, particularly in black-box settings where the internal workings of the model generating the text are not accessible. Even in cases where the underlying model is known, such as in the SHROOM competition, we adhere to a strictly black-box approach, meaning that we do not regenerate outputs from the models but instead focus solely on the detection of hallucinations in the existing outputs<sup>1</sup>.

---

<sup>1</sup>Even in the model-aware setting of SHROOM, we do not re-generate the outputs using the given models, therefore we continue operating in a completely black-box setup.

To this end, we leverage pre-trained models that incorporate knowledge of semantic relationships relevant to hallucinations. Through fine-tuning these models on labeled instances from the SHROOM dataset, we achieve an approximate accuracy of 80% in detecting hallucinations, which significantly surpasses the baseline accuracy of 74.5% achieved using an open-source Mistral instruction-tuned model<sup>2</sup>.

This work makes several key contributions:

1. We fine-tune models pre-trained on hallucination detection and Natural Language Inference (NLI) datasets, which are semantically aligned with the challenges presented by SHROOM.
2. The fine-tuned models are integrated into a Voting Classifier framework, which yields competitive detection accuracy in the SHROOM tasks.
3. We design a novel prompting method that enhances the interpretability of hallucination detection by guiding models to focus on key semantic inconsistencies in generated outputs.
4. Sentence similarity metrics, derived from embeddings produced by pre-trained models, are leveraged to quantify and detect hallucinations by comparing the generated outputs to the input data.
5. All experimentation conducted in this work is designed to be time and computationally efficient, adhering to the constraints of a fully black-box environment.
6. A detailed analysis of the results, broken down by task, is provided, offering insights into both the successful detection of hallucinations and the challenges posed by failed instances, contributing to a deeper understanding of the nature of hallucinations in LLM outputs.

In sum, this research not only advances the state of the art in hallucination detection for LLMs but also provides a foundation for future work aimed at mitigating this inherent issue in natural language generation tasks. By exploring the intersection of model architecture, task specificity, and evaluation metrics, we contribute to the development of more reliable and trustworthy LLM-based systems.

---

<sup>2</sup><https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF>

# Chapter 3

## Related Work

### Contents

|                                            |           |
|--------------------------------------------|-----------|
| <b>3.1 Large Language Models</b> . . . . . | <b>49</b> |
| <b>3.2 Transformers</b> . . . . .          | <b>50</b> |
| <b>3.3 LLMs Hallucinations</b> . . . . .   | <b>51</b> |

### 3.1 Large Language Models

*Language Models (LMs)* are computational frameworks designed to understand and generate human language. More precisely, a language model defines a probability distribution over sequences of words, enabling the prediction of the likelihood of these sequences or the generation of new text based on an input.

Historically, the fundamental approach for probabilistic language modeling has been *n-gram models*. These models follow the Markov chain assumption, where the probability of the next word in a sequence is conditioned on a fixed window of preceding words. For example, a bigram model considers one preceding word, a trigram two, and an n-gram model considers  $n - 1$  preceding words. The probability of a sequence  $w_1, w_2, \dots, w_n$  in a bigram model is represented as follows:

$$P(w_1, w_2, \dots, w_n) = P(w_2|w_1) \cdot P(w_3|w_2) \cdots P(w_n|w_{n-1}) \quad (3.1.1)$$

Here, the conditional probability  $P(w_k|w_{k-1})$  is estimated by the relative frequency of occurrences where  $w_{k-1}$  is followed by  $w_k$  in the training corpus. Training an n-gram model involves estimating these probabilities from text corpora, typically in one or more languages.

A key limitation of n-gram models is their inability to assign non-zero probabilities to sequences not encountered during training, given the infinite range of valid sentences a language can express. Various smoothing techniques have been proposed to address this issue.

The advent of neural networks introduced the era of neural language models, starting with simple feed-forward neural language models [4]. A feed-forward neural language model is a straightforward neural network that takes a fixed number of preceding words as input and outputs a probability distribution over possible next words. Similar to an n-gram model, a feed-forward neural language model predicts the probability of a word conditioned on the  $n - 1$  preceding words.

A major distinction between n-gram models and feed-forward neural language models lies in the representation of word sequences. While n-gram models use discrete word identities, feed-forward neural language models leverage word embeddings, continuous vector representations of words. This results in superior generalization performance, particularly for unseen word sequences during testing.

Although feed-forward neural language models marked a significant advancement in language modeling, modern neural language models are built on more sophisticated architectures such as recurrent networks and transformers.

*Recurrent neural network (RNN)* language models [36] process input sequences sequentially, analyzing one word at a time. Their goal is to predict the next word based on the current word and the previous hidden state. Unlike n-gram models, RNNs can capture long-range dependencies in the input sequence, as they do not rely on a fixed-size context. The hidden state in RNNs can, in theory, represent information from all preceding words in the sequence, addressing the limited context of n-gram models and feed-forward neural networks.

## 3.2 Transformers

*Large Language Models (LLMs)* represent sophisticated language models characterized by vast numbers of parameters and exceptional learning capacities, typically pre-trained on extensive unstructured text corpora. Modern LLMs are built predominantly on the Transformer architecture [48], which eliminates the need for recurrence and instead employs a self-attention mechanism to capture global dependencies between inputs and outputs. The Transformer consists of two primary components: the *encoder* and *decoder*. These components are essential for sequence-to-sequence tasks like machine translation, where the encoder processes the input sequence and the decoder generates the corresponding output sequence.

The core components of the Transformer architecture are summarized below:

1. **Input Representation:** The input sequence is initially embedded into continuous vector representations. Positional embeddings are then added to these vectors to encode the position of each token in the sequence.
2. **Encoder:** The encoder is composed of a stack of identical layers, each containing two sub-layers:
  - **Multi-Head Self-Attention:** This sub-layer computes a weighted sum of embeddings, enabling each token to focus on different parts of the input sequence. Multiple attention heads operate in parallel, capturing various relationships between tokens.
  - **Position-wise Feed-Forward Neural Network:** After the attention mechanism, each token's representation is passed through a position-wise feed-forward neural network. This layer introduces non-linearity and further refines the token embeddings.

Residual connections [18], followed by layer normalization [3], are applied around each sub-layer.

3. **Decoder:** The decoder also consists of a stack of identical layers, but with three sub-layers:
  - **Masked Multi-Head Self-Attention:** Similar to the encoder's self-attention mechanism, this sub-layer includes a mask to prevent the model from attending to future positions during training.
  - **Multi-Head Encoder-Decoder Attention:** This sub-layer attends to the encoder's output, allowing the decoder to focus on relevant parts of the input sequence during generation.
  - **Position-wise Feed-Forward Neural Network:** Like the encoder, this sub-layer further refines the token representations after attention.

As with the encoder, residual connections are applied around each sub-layer, followed by layer normalization.

4. **Output Generation:** The output of the final decoder layer is transformed into probability distributions over the output vocabulary using a linear transformation followed by a softmax activation function. During training, the model is provided with an input sequence to predict the next word.

Since its introduction [48], several variations of the Transformer architecture have been proposed. A key distinction among these variations lies in the masking pattern applied to the inputs, which provides the contextual information needed for the model to generate accurate predictions.

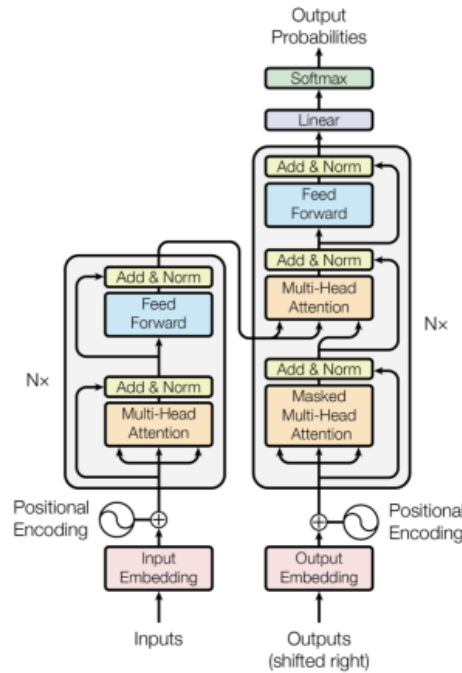


Figure 3.2.1: Transformer Architecture

**Encoder-Decoder:** As initially proposed, the Transformer consisted of two components: the encoder and decoder. The encoder processes the input sequence, creating context-rich representations that the decoder uses to generate the output sequence step-by-step. Prominent pre-trained models using the encoder-decoder architecture include BART [25] and T5 [42].

**Decoder-Only:** Although the encoder-decoder structure was the original Transformer variation, modern LLMs primarily use a decoder-only architecture. In this design, models are trained as standard language models, learning to predict the next token in a sequence. Unlike the encoder-decoder architecture, decoder-only models do not process input and output sequences separately; all tokens are processed equally, with conditioning based only on prior tokens due to the causal masking pattern. This design is simpler and ideal for autoregressive next-token prediction tasks. Notable LLMs using this architecture include the GPT series [5, 41] and several other recent models [50],[47].

**Encoder-Only:** Additionally, an encoder-only variant of the Transformer exists, where only the encoder stack is employed. This architecture forms the foundation of models such as BERT [10] and its derivatives, which are designed primarily for tasks like text classification and sentence representation.

Overall, Transformers have revolutionized NLP by efficiently handling sequential data, enabling parallel processing, and capturing long-range dependencies within texts. The attention mechanism allows Transformers to model relationships between input and output data without the need for convolutions or recurrent units, achieving state-of-the-art performance on numerous language processing tasks.

### 3.3 LLMs Hallucinations

NLP hallucinations have emerged as a critical challenge in the development of natural language generation (NLG) systems, and research in this field has been rapidly evolving, exploring hallucinations from diverse perspectives. Broadly speaking, hallucinations in NLP can be categorized based on whether the generated outputs are unfaithful to the input, inconsistent within the generated text, or incongruent with real-world knowledge [53]. Among these, factual hallucinations—where the model generates information not supported by external factual sources—have gained considerable attention. This is largely because factual hallucinations

are easier to detect and correct through comparison with existing databases or knowledge repositories [29, 24, 7, 37, 6, 38]. Several recent studies have made significant progress in building frameworks and tools to address this issue, leading to more reliable and accurate detection mechanisms [8, 11, 28].

Despite these advancements in factual hallucination detection, other forms of hallucinations—such as semantic and fluency-based hallucinations—remain underexplored, making it difficult to develop benchmarks and automated evaluation techniques that cover all types of hallucinatory outputs. Semantic hallucinations, for instance, involve cases where generated content is internally coherent but semantically incorrect or misaligned with the input. The inherent subtlety of such hallucinations makes the construction of comprehensive benchmarks and evaluation systems more challenging [2, 23, 33, 12]. Moreover, many of the current techniques for hallucination detection depend on probing the model itself, which becomes impractical in scenarios where the model is proprietary, closed, or inaccessible. This limitation is a significant barrier to the broader application of these methods in real-world scenarios, where the system generating the outputs may be opaque or entirely unavailable to the researcher.

In response to this gap, the SHROOM framework has been introduced, focusing primarily on the concept of semantic faithfulness rather than factual correctness. SHROOM diverges from traditional methods by advocating for detection techniques that do not require access to the model that generated the outputs. Instead, it evaluates hallucinations based on the provided input and output alone. This design makes SHROOM applicable even when the underlying NLG model is inaccessible, as is often the case with closed commercial systems. However, a trade-off in SHROOM’s approach is its reliance on ground-truth outputs, which must be available for meaningful evaluation, since the input alone may not provide sufficient semantic context to determine whether the output is hallucinatory. This approach differs from other black-box methodologies that rely on model probing [33], further underscoring the importance of adaptable hallucination detection methods that can operate across different settings.

The recognition that hallucinations frequently occur in NLG models has been well-established, with early work by Vinyals and Le [49] and later studies such as Raunak et al. [43] and Maynez et al. [34] highlighting this phenomenon across various domains. Nevertheless, there is still minimal consensus on the optimal framework for detecting and addressing hallucinations in NLG systems. This lack of agreement can be attributed to the diversity of tasks that NLG encompasses, ranging from machine translation and summarization to dialogue systems, each of which may exhibit different types of hallucinations [21]. As a result, the community has yet to converge on a standardized approach that can address hallucinations across multiple tasks.

Guerreiro et al. [17] have attempted to address this issue by proposing a taxonomy of hallucinations that categorizes different types of hallucinatory outputs, such as oscillatory productions—where a model generates repetitive or redundant text—and fluent but detached outputs, which, though linguistically coherent, are semantically disconnected from the input. While this taxonomy represents an important step toward better understanding hallucinations, it has certain limitations. For example, the taxonomy conflates fluency with semantic correctness, especially in cases where oscillatory productions are involved, and focuses primarily on extreme cases of hallucinations, such as fully detached outputs. Yet, intermediary hallucination cases—those that are not as obvious but are more challenging to detect—are often more frequent and have greater relevance for practical applications. Moreover, the taxonomy primarily addresses machine translation (MT), whereas hallucinations also occur in other NLG tasks, including text summarization and open-domain dialogue systems [46].

Several studies have developed benchmarks for hallucination detection, each catering to different aspects of NLG. Li et al. [27] introduced the HaluEval benchmark, which employs an annotation framework focused on fact-checking but does not necessarily center on the model’s input. While this benchmark provides valuable insights, it has certain limitations: annotators are required to search the internet for supporting evidence, which can make the annotation process cumbersome. Furthermore, HaluEval annotates outputs from a popular but closed and commercial large language model (LLM), limiting the transparency and generalizability of findings from future research based on the benchmark. Other benchmarks, such as those by Liu et al. [30] and Zhou et al. [54], introduce synthetic hallucinations by manipulating training data for token-level detection, while Lin et al. [29] focus on factual claims supported by publicly available knowledge bases. Dziri et al. [13], on the other hand, emphasize knowledge-based dialogue systems, using natural language inference (NLI) to identify hallucinations without relying on the model itself.

Our work builds on this body of research, particularly aligning with approaches like that of Dziri et al. [13], which rely solely on the input-output relationship rather than probing the underlying model. This is crucial, as it allows for the detection of hallucinations in scenarios where the model is inaccessible or proprietary, thus broadening the applicability of hallucination detection methods. While many existing benchmarks and tools focus heavily on factual hallucinations or rely on extensive fact-checking protocols, our approach places greater emphasis on semantic faithfulness, aiming to develop robust methods that can handle a wide range of NLG tasks, irrespective of model transparency.



# Chapter 4

## Approach

### Contents

---

|            |                                         |           |
|------------|-----------------------------------------|-----------|
| <b>4.1</b> | <b>Dataset</b>                          | <b>55</b> |
| 4.1.1      | Data format                             | 55        |
| 4.1.2      | Exploratory data analysis               | 56        |
| <b>4.2</b> | <b>Metrics</b>                          | <b>64</b> |
| 4.2.1      | Accuracy                                | 64        |
| 4.2.2      | Spearman Correlation                    | 64        |
| <b>4.3</b> | <b>Methods</b>                          | <b>65</b> |
| 4.3.1      | Sentence Similarity                     | 65        |
| 4.3.2      | COMET                                   | 65        |
| 4.3.3      | Fine-tune NLI models                    | 66        |
| 4.3.4      | Fine-tune hallucination detection model | 67        |
| 4.3.5      | Voting Classifier                       | 68        |
| 4.3.6      | Prompting                               | 68        |

---

### 4.1 Dataset

Driven by upcoming challenges in the NLG landscape, SHROOM dataset focuses on the prevalent issues of models generating linguistically fluent but inaccurate (incorrect or unsupported) outputs. Participants were tasked with binary classification to identify instances of fluent overgeneration hallucinations in *model-aware* and *model-agnostic* tracks. The task encompasses three NLG domains—definition modeling (DM), machine translation (MT), and paraphrase generation (PG)—with provided checkpoints, inputs, references, and outputs for binary classification. The development set includes annotations from multiple annotators, establishing a majority vote gold label.

#### 4.1.1 Data format

In all cases, data follow a specific format:

- *src* is the input given to a model, *hyp* is the output generated by the model
- *tgt* comprises the ground truth output for this specific model
- *ref* indicates whether target, source or both of these fields contain the semantic information necessary to establish whether a datapoint is a hallucination,
- *task* refers to the task being solved and
- *model* to the model being used (in the model-agnostic case the *model* entry remains empty).

Initially, 80 labeled trial samples were released, followed by unlabelled training data which contain 30k model-agnostic and 30k model-aware instances. Finally, the labeled validation set contains 499 and 501 samples for model-agnostic and model-aware settings respectively, while the test set comprises 1500 model-agnostic and 1500 model-aware labeled samples. Additional information provided in the labeled splits are *labels*, which contains a list of ‘Hallucination’ and ‘Not Hallucination’ labels as provided by 5 annotators per sample, the final *label* occurring via majority voting over the aforementioned list and  $p(\text{Hallucination})$ , denoting the probability of hallucination as the percentage of agreeing annotators on the ‘Hallucination’ label.

During the first phase of the competition, only the validation set was released, hence the experiments were conducted on the validation set, as opposed to the second phase, when the final experiments were performed on the test set.

In Table 4.1 we present some examples from the unlabelled training dataset containing model-agnostic and model-aware instances. Regarding the machine translation (MT) task, we could detect a variety of languages, including Russian, Arabic, Chinese, Yorùbá, Telugu, Tsonga, Uzbek, Sinhalese, Quechuan, Mizo and others. Language information was not provided, so we manually explored the *src* samples in terms of linguistic variability.

Model-agnostic definition modeling (DM) hypotheses contain some ‘qualifiers’, which may guide a model under usage to return a more suitable definition. For example, in the context of the hypothesis containing the definition "(obsolete) An odour," the term "obsolete" indicates that the provided definition is no longer in common use or is outdated. The word "obsolete" is used as a qualifier to convey that the term or concept being defined, in this case, "An odour," was once used to represent a specific meaning but is no longer considered current or applicable in contemporary language.

Another notable observation is that model-aware paraphrase-generation (PG) contains no information in *tgt*.

### 4.1.2 Exploratory data analysis

**Trial set** We explore the frequency of each task occurring within samples from different dataset splits, commencing from the initially released trial set. In Figure 4.1.1 we present the task distribution of the first 80 trial samples.

**Unlabelled data (training set)** Figure 4.1.2 represents the distribution in the training set. In both model-agnostic and model-aware settings each task contains an equal number of samples (10k samples per task in each setting). In our methodologies, we abstained from utilizing the provided unlabeled training dataset as it did not align with our main approaches.

**Validation set** Moving on to labeled data, we commence with the validation (dev) set, for which we present per task distributions in Figure 4.1.3. We observe a difference in the distribution of labels in comparison to the balanced training set distribution of Figure 4.1.2; nevertheless, since we do not exploit any unlabelled instance, this does not pose a limitation for us at this point.

We proceed with studying the validation set label distribution. Related results are presented in Figure 4.1.4, denoting label imbalance in both model-agnostic and model-aware settings.

The distribution of hallucination probability is presented in Figure 4.1.5. As expected, low  $p(\text{'Hallucination'})$  values are more common (indicating that fewer annotations voted for the presence of a hallucinatory instance), since ‘Not Hallucination’ is the majority label in both settings. Ideally, we wish borderline probabilities to be low: The highest the disagreement for a certain sample, the closest to the 0.5 threshold the hallucination probability will be (a  $p(\text{'Hallucination'})=0.4$  denotes that 3/5 annotators voted for ‘Not Hallucination’, while the rest 2/5 voted for the opposite; on the other hand, a  $p(\text{'Hallucination'})=0.6$  denotes that 3/5 annotators voted for ‘Hallucination’, while the rest 2/5 voted for ‘Not Hallucination’). Therefore, the highest uncertainty is observed close to the 0.5 boundary). This requirement is adequately satisfied especially in the model-agnostic case (left plot of Figure 4.1.5), where  $p(\text{'Hallucination'})=0.6$  is the least frequent.

Further insights can be obtained by looking at Figure 4.1.6: when smaller dots are assigned to probabilities close to the 0.5 threshold, the annotators’ disagreement is lower, therefore classifying a sample as ‘Halluci-

| Model-agnostic        |                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Machine Translation   | 'hyp': "Don't worry, it's only temporary.", 'tgt': "Don't worry. It's only temporary.", 'src': 'Не волнуйся. Это только временно.', 'ref': 'either', 'task': 'MT', 'model': "                                                                                                                                                                                                                                                                              |
| Definition modelling  | 'hyp': '(uncountable) The quality of being oronymy; the state of being oronymy.', 'tgt': 'The nomenclature of mountains, hills and other geographic rises.', 'src': 'An ancient survival in Turkish <define> oronymy </define> is quite possible , but I have not found Nihan Dag on the relevant sheets of the 1 : 200,000 map of Turkey , which are very detailed in matters of oronymy ;', 'ref': 'tgt', 'task': 'DM', 'model': "                       |
| Definition modelling  | 'hyp': '(intransitive, obsolete) To make a magazin of; to compose a magazin.', 'tgt': '(colloquial) The act of editing or writing for a magazine.', 'src': "Thus , though Byron is gone after his Don Juan — Scott and Southey out of the rhyme department — Wordsworth stamp - mastering — Coleridge 's poetry in abeyance — Crabbe mute as a fish - Campbell and Wilsont merely <define> magazinizing </define>", 'ref': 'tgt', 'task': 'DM', 'model': " |
| Paraphrase Generation | 'hyp': 'You got something for me, huh?', 'tgt': ", 'src': 'Got something for me?', 'ref': 'src', 'task': 'PG', 'model': "                                                                                                                                                                                                                                                                                                                                  |
| Model-aware           |                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Machine Translation   | 'hyp': 'Our Mailoamiris of the System of Treatment of Ulilaes have created a place for these little ones.', 'tgt': 'We perceive the Foster Care System to be a safety zone for these children.', 'src': 'Maamiris tayo a ti Sistema iti Panangtaripato kadagiti Ulila ket natalged a lugar para kadagitoy nga ubbing.', 'ref': 'either', 'task': 'MT', 'model': 'facebook/nllb-200-distilled-600M'                                                         |
| Definition modeling   | 'hyp': 'To be obsequiously interested in .', 'tgt': '( usually followed by over or after ) To fuss over something adoringly ; to be infatuated with someone .', 'src': "Sarah mooned over sam 's photograph for months . What is the meaning of moon ?", 'ref': 'tgt', 'task': 'DM', 'model': 'ltg/flan-t5-definition-en-base'                                                                                                                             |
| Paraphrase Generation | 'hyp': "Mr Barros Moura's report looks to the future in my opinion.", 'tgt': ", 'src': 'In my opinion, the most important element of the report by Mr Barros Moura is that it looks to the future.', 'ref': 'src', 'task': 'PG', 'model': 'tuner007/pegasus_paraphrase'                                                                                                                                                                                    |

Table 4.1: Examples from the unlabelled training set.

nation' or 'Not hallucination' is less uncertain. Indeed, the less frequently appearing  $p(\text{'Hallucination'})=0.4$  and  $p(\text{'Hallucination'})=0.6$  values in the model-agnostic case denote high separability between hallucinated and non-hallucinated samples. However, highly certain values, such as  $p(\text{'Hallucination'})=0.0$  and  $p(\text{'Hallucination'})=1.0$  only rank in the middle, therefore even if samples are separable with low uncertainty, some minor disagreement persists (1/5 annotators frequently disagrees with the rest). Overall, annotators are almost equally confident in classifying 'Hallucination' and 'Not Hallucination' samples, as indicated by the matching pattern regarding label uncertainty for both labels. The model-aware case is more confusing, with  $p(\text{'Hallucination'})=0.6$  scoring the highest; therefore, classifying a sample as 'Hallucination' is often accompanied by high uncertainty. On the contrary, uncertainty is lower for the 'Not Hallucination' label, with  $p(\text{'Hallucination'})=0.0$  ranking as the second most frequent probability. We can conclude that in the model-aware setting of the validation set, annotators are more confident in recognizing the 'Not Hallucination' class in comparison to the 'Hallucination' one.

**Test set** As for the test set, Figure 4.1.7 represents the number of samples per task for both settings. Note that the test task distribution is similar to the validation distribution of Figure 4.1.3 with PG being a minority label in all cases.

In terms of ground-truth label (Hallucination vs Not Hallucination), Figure 4.1.8 highlights some label imbalance, rendering the prediction of 'Not Hallucination' more possible in a random setup for both model-agnostic and model-aware settings. This label distribution matches the validation set label distribution (Figure 4.1.4), for which 'Not Hallucination' was the majority class as well.

Hallucination probability per setting is depicted in Figure 4.1.9, with lower hallucination values in the range  $[0, 0.2)$  being more common. This is again somehow expected since 'Not Hallucination' is the majority class in test labels. More insights can be obtained by looking at Figure 4.1.10, which relates the hallucination probabil-

## Number of samples per task (trial data)

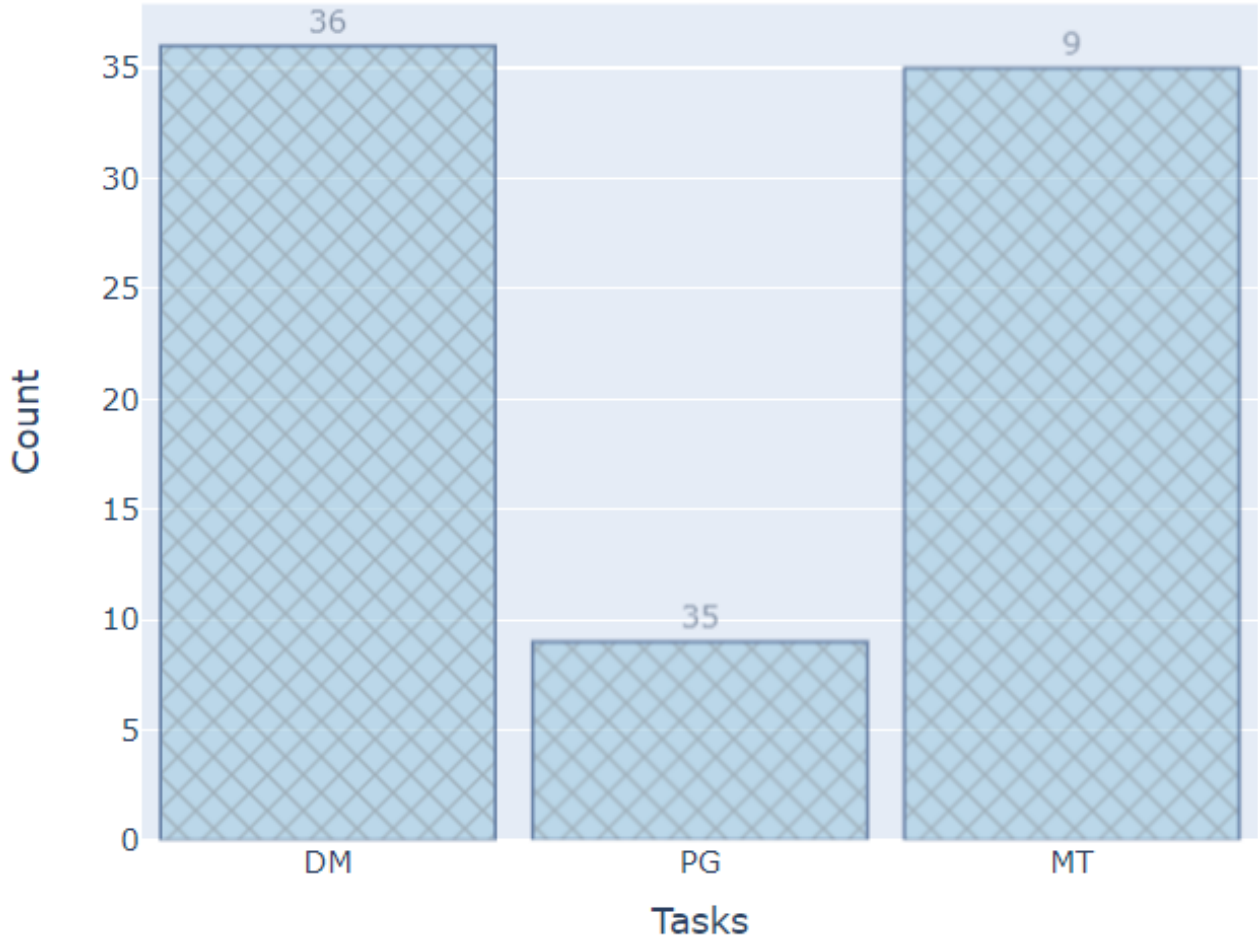


Figure 4.1.1: Distribution of per task samples in the initially released trial set.

ity with the label. Especially in the model-agnostic setting (Figure 4.1.10 - left), the  $p(\text{'Hallucination'})=0.4$  and  $p(\text{'Hallucination'})=0.6$  values are the lowest (smaller dots), while  $p(\text{'Hallucination'})=0.0$  is the highest, denoting that annotators are often certain regarding non-hallucinated samples. Certainty for hallucinated samples is somehow lower, as  $p(\text{'Hallucination'})=1.0$  lies somewhere in the middle. Nevertheless,  $p(\text{'Hallucination'})=0.8$  is the second more frequent value denoting that 4/5 annotators frequently annotate a sample as 'Hallucination'. By observing the right plot of Figure 4.1.10, we conclude that certainty is lower in the model-aware setting. Even though  $p(\text{'Hallucination'})=0.0$  remains the most frequent probability, indicating high agreement regarding non-hallucinated samples, the  $p(\text{'Hallucination'})=0.6$  value stands in the second place. Therefore, many samples classified as 'Hallucination' achieved this label with low agreement (3/5 annotators). Also, the  $p(\text{'Hallucination'})=0.2$  and  $p(\text{'Hallucination'})=0.8$  are the lowest, denoting that higher agreement (4/5 annotators agreeing) is rare for both 'Hallucination' and 'Not Hallucination' labels. We can assume that model-aware samples are harder by nature to be classified in any of the labels.

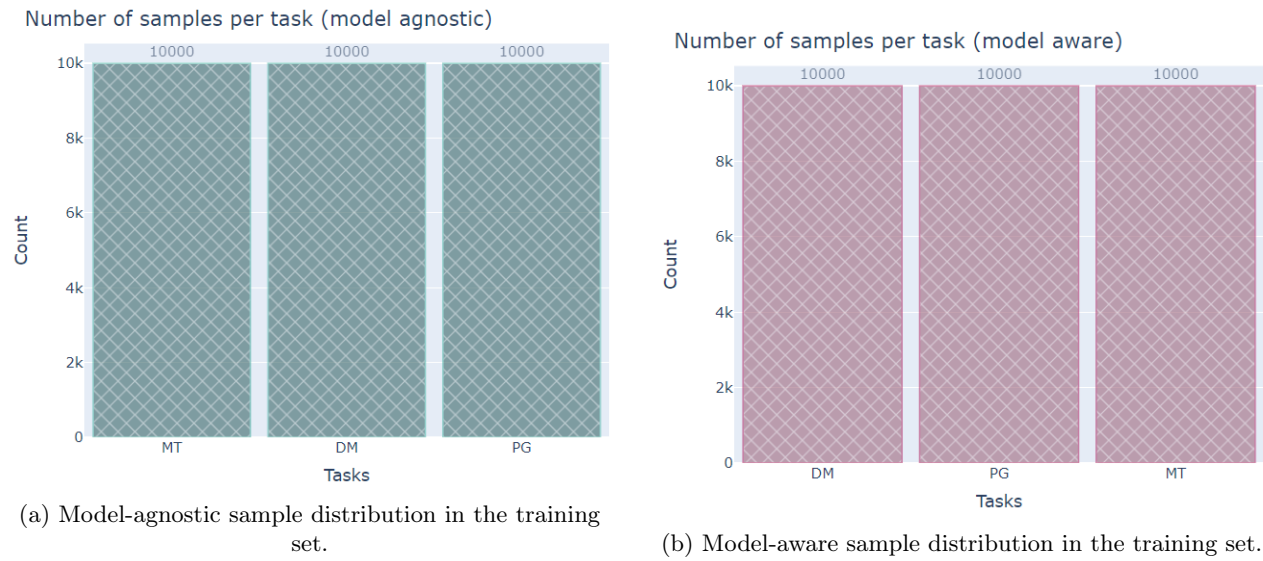


Figure 4.1.2: Distribution of unlabelled training samples per task in both model-agnostic and model-aware settings.

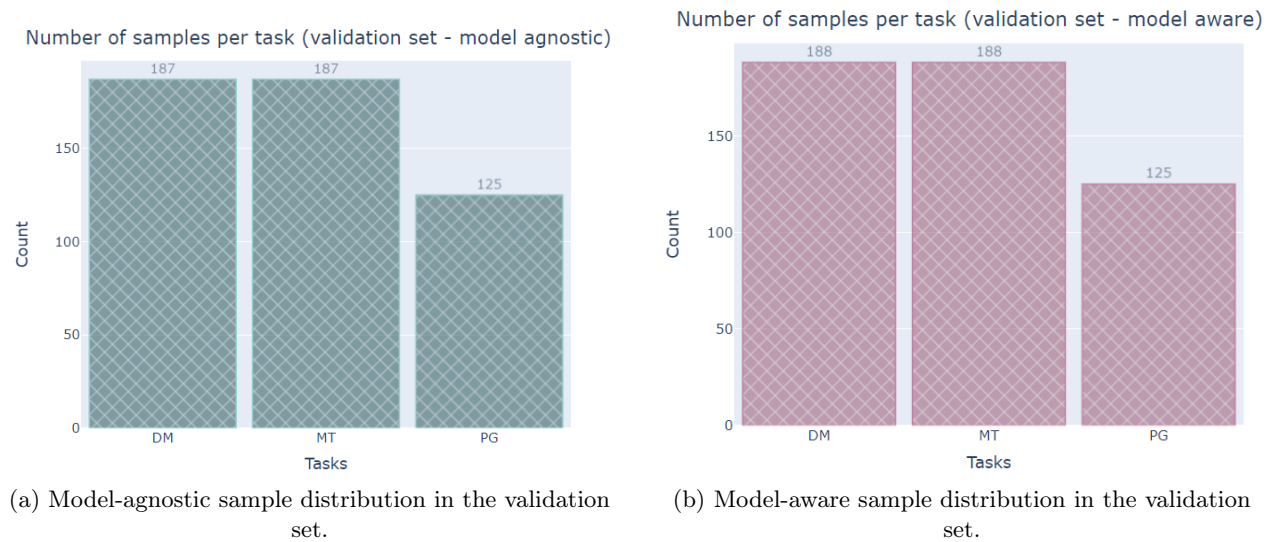
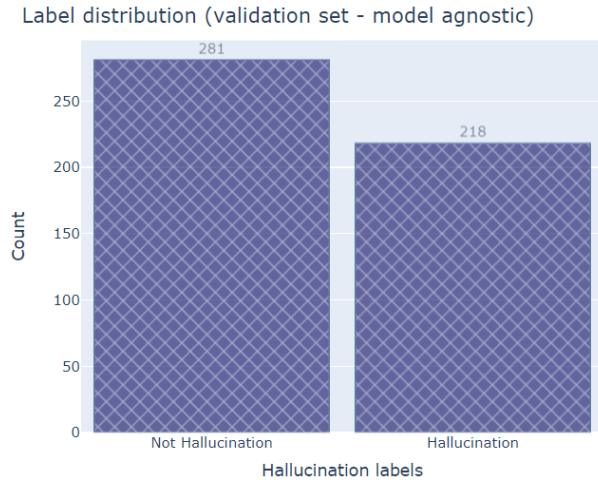
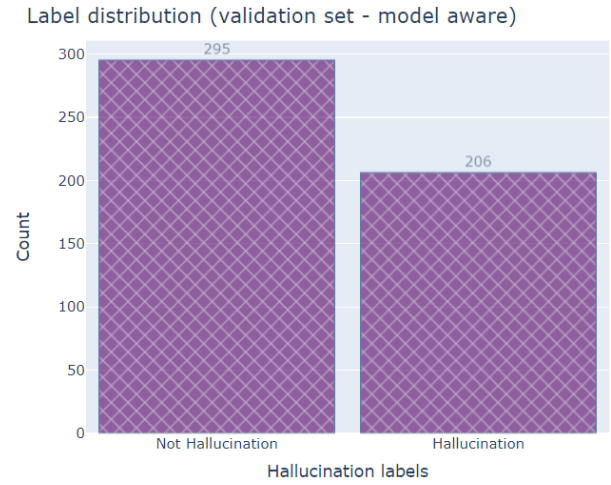


Figure 4.1.3: Distribution of labeled validation samples per task in both model-agnostic and model-aware settings.

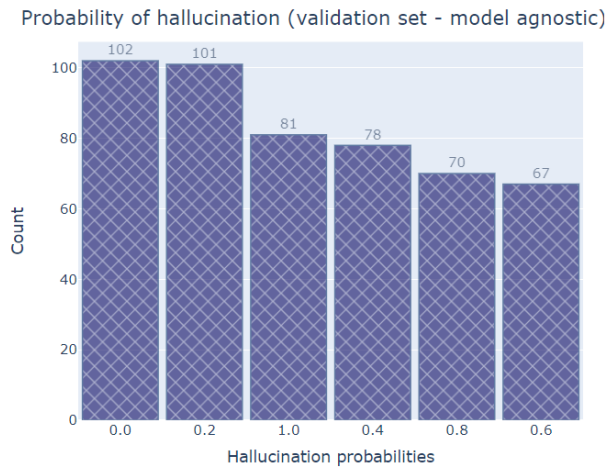


(a) Model-agnostic label distribution in the validation set.

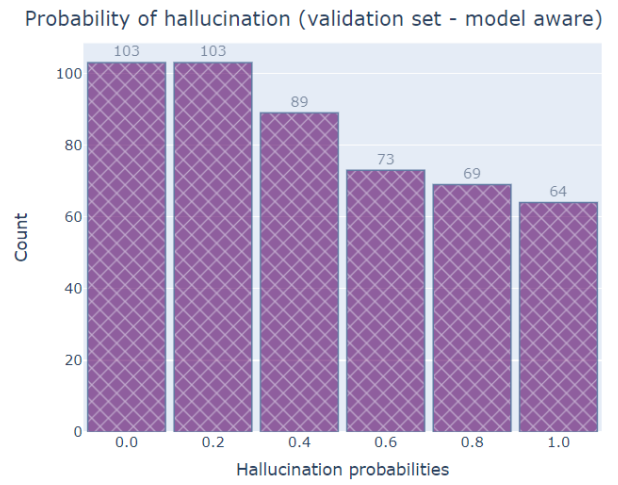


(b) Model-aware label distribution in the validation set.

Figure 4.1.4: Distribution of validation labels in both model-agnostic and model-aware settings.



(a) Model-agnostic hallucination probability distribution in the validation set.



(b) Model-aware hallucination probability distribution in the validation set.

Figure 4.1.5: Distribution of hallucination probability (majority voting among human annotators' labeling) in both model-agnostic and model-aware settings in the validation set.

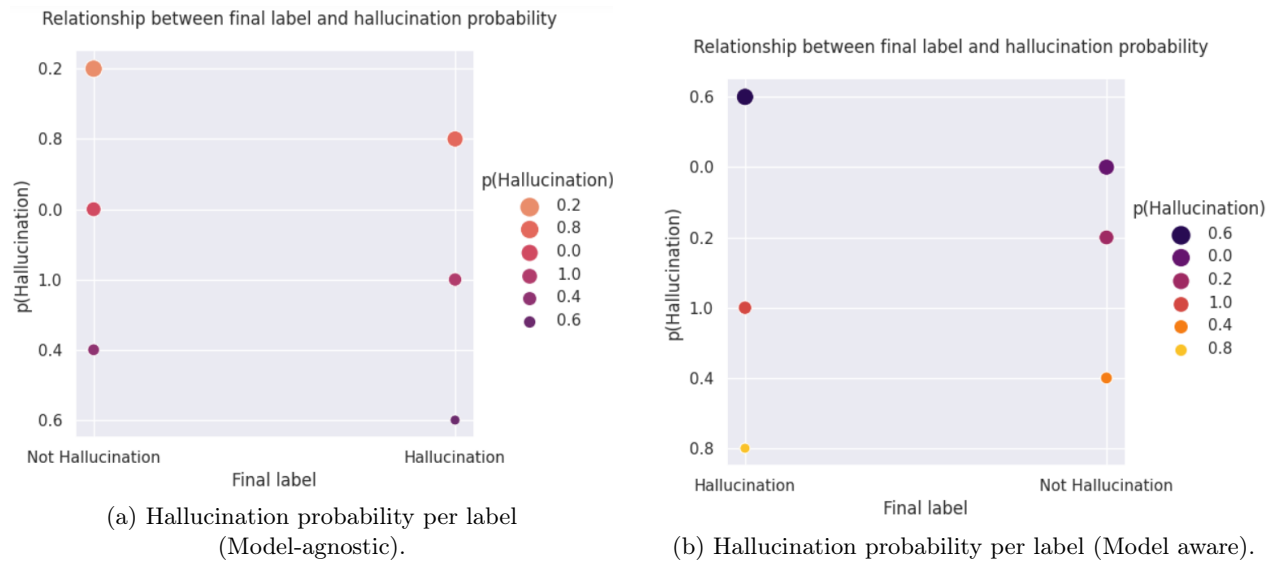


Figure 4.1.6: Distribution of hallucination probability in each validation label ('Hallucination' vs 'Not Hallucination'). Annotators significantly agree on whether a sample contains a hallucination or not.

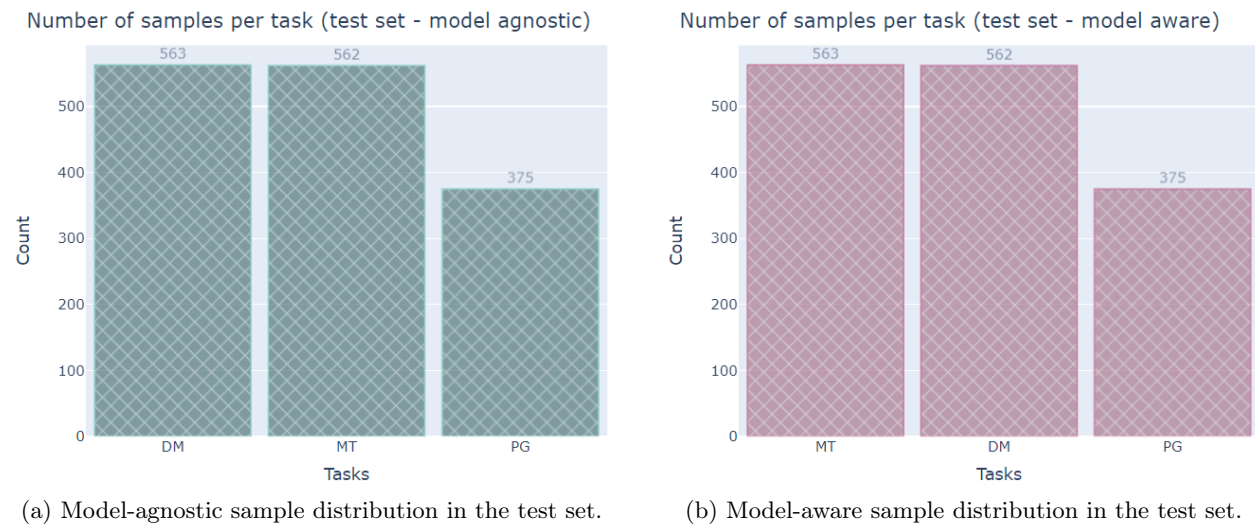


Figure 4.1.7: Distribution of labeled test samples per task in both model-agnostic and model-aware settings.

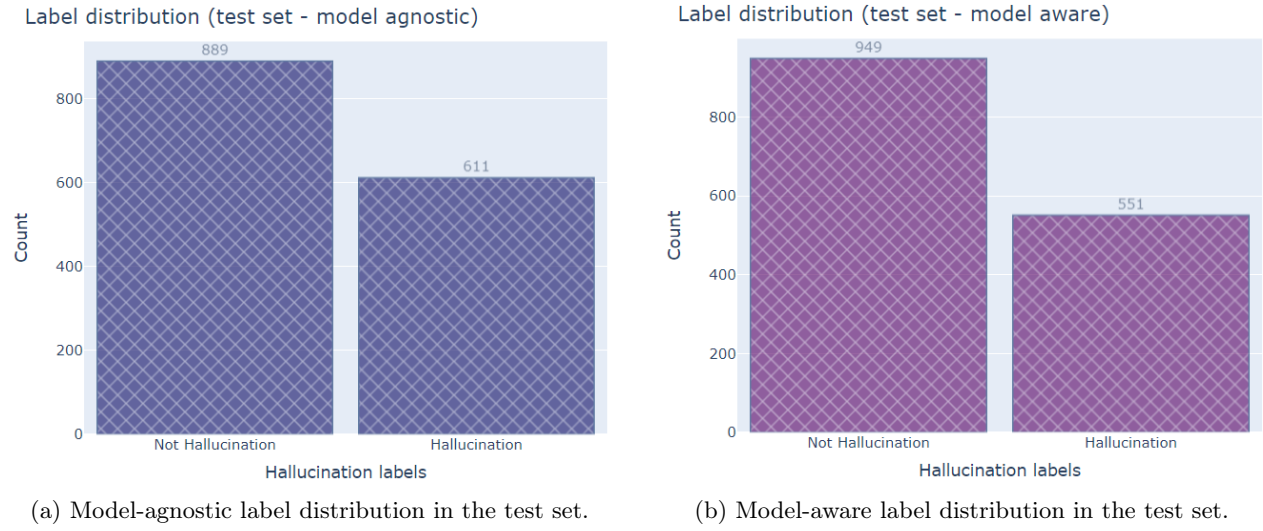


Figure 4.1.8: Distribution of test labels in both model-agnostic and model-aware settings.

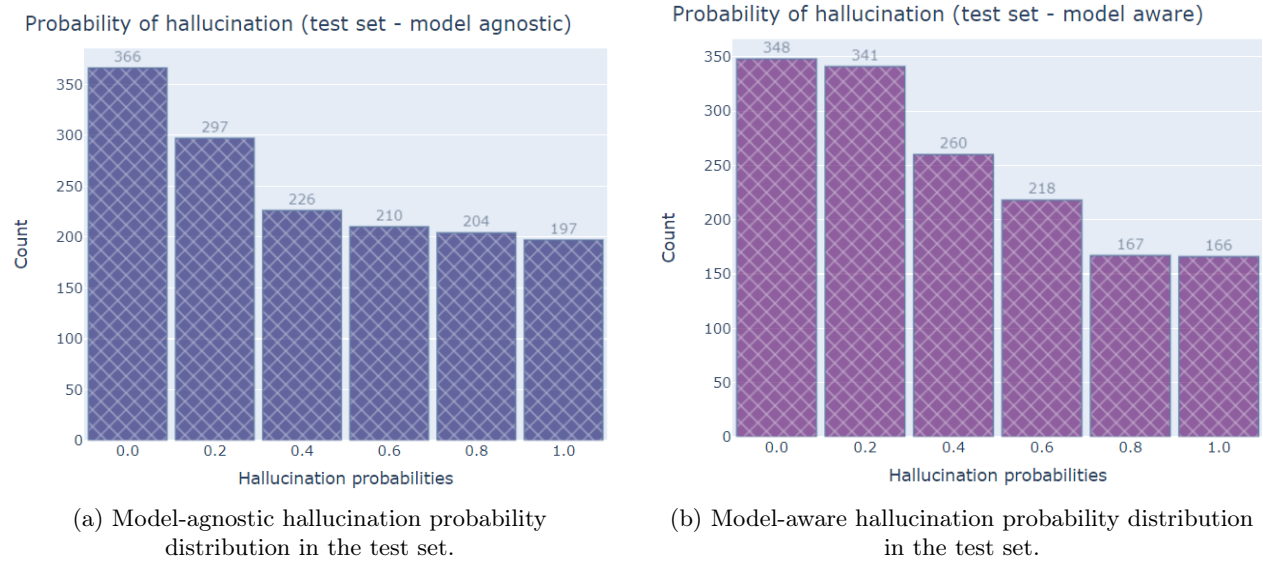


Figure 4.1.9: Distribution of hallucination probability (majority voting among human annotators' labeling) in both model-agnostic and model-aware settings in the test set.

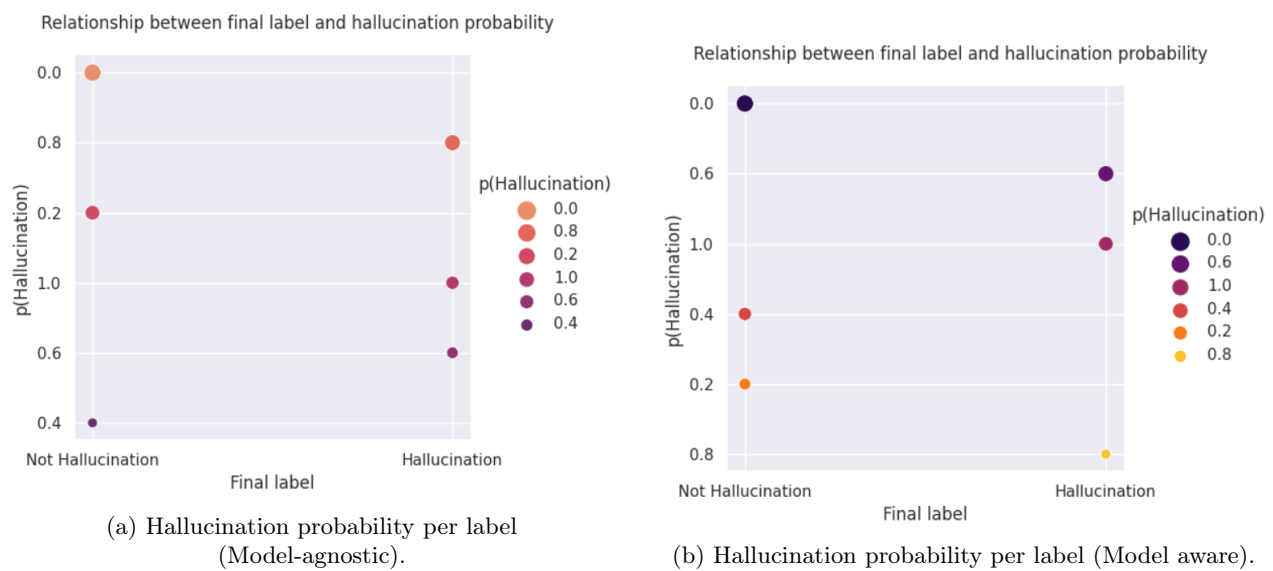


Figure 4.1.10: Distribution of hallucination probability in each test label ('Hallucination' vs 'Not Hallucination'). Annotators significantly agree on whether a sample contains a hallucination or not.

## 4.2 Metrics

In order to evaluate the results of the experiments we used two metrics, according to the evaluation method of the Semeval completion: Accuracy for the binary labels ('Hallucination' and 'Not Hallucination') and Spearman correlation for the percentage of Hallucination ('p(Hallucination)').

### 4.2.1 Accuracy

Accuracy[1] is a metric used to determine how frequently a machine learning model makes correct predictions. It is calculated by dividing the number of correct predictions by the total number of predictions made, as shown in the following equation:

$$Accuracy = \frac{CorrectPredictions}{AllPredictions} \quad (4.2.1)$$

In simple terms, accuracy tells us how often the model gets it right.

### 4.2.2 Spearman Correlation

The Spearman correlation[9] is a statistical measure that assesses the strength and direction of a monotonic relationship between two variables by comparing their ranked values. For instance, a data series such as [58, 70, 40] would be ranked as [2, 1, 3]. The Spearman correlation is particularly useful when dealing with data containing outliers. To quantify this correlation, the Spearman rank correlation coefficient  $r_s$ , also referred to as the rank coefficient, is employed. The formula for  $r_s$  is given below, where  $n$  represents the number of data points in the series, and  $d$  denotes the difference in ranks between corresponding values of the two variables:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (4.2.2)$$

Here,  $d_i$  is the difference between the ranks of the two corresponding values in the data pairs.

Systems were evaluated according to two criteria: the accuracy that the system reached on the binary classification, and their calibration, measured as the Spearman correlation of the systems' output probabilities with the proportion of the annotators marking the item as overgenerating. The systems were ranked by accuracy and break possible ties using spearman correlation.

## 4.3 Methods

### 4.3.1 Sentence Similarity

The initial approach to solving the problem involved reframing it as a task of sentence similarity. Sentence similarity refers to the process of measuring the degree of similarity between two textual inputs based on their semantic content. This process is central to many natural language processing (NLP) tasks, as it enables the comparison of sentences not merely on a lexical level but through their underlying meaning. To perform this task, models typically represent the input texts as vectors, or embeddings, that encode their semantic information. The similarity between these embeddings is then quantified by calculating their proximity within a high-dimensional vector space.

In our implementation of this method, we took several steps to adapt the sentence similarity task to our specific problem. First, we encoded both the target ('tgt') and hypothesis ('hyp') texts for each datapoint. To do this, we utilized various pre-trained models from the `Sentence_transformers` library, which is known for generating high-quality sentence embeddings that capture semantic nuances effectively. These embeddings serve as vector representations of the sentences, preserving their meaning while allowing for mathematical manipulation.

After obtaining the embeddings, we computed the distance between the 'tgt' and 'hyp' vectors. To ensure a comprehensive evaluation of similarity, we employed multiple distance metrics, specifically the Manhattan distance and the Euclidean distance. These metrics allowed us to quantify the extent of difference or similarity between the sentence pairs by measuring how far apart their corresponding vectors are in the embedding space.

Finally, based on the calculated distances, we established an appropriate threshold for binary classification. This threshold was determined by experimenting with various distance values, allowing us to define a cutoff point that effectively distinguishes between similar and dissimilar sentence pairs. Through this process, we transformed the sentence similarity task into a binary classification problem, where the similarity score was used to classify each sentence pair as either similar or dissimilar based on the predefined threshold.

This method provided an initial framework for tackling the problem, leveraging well-established techniques in sentence similarity to compare the semantic content of textual inputs and determine the necessary classification.

### 4.3.2 COMET

COMET [45], (Crosslingual Optimized Metric for Evaluation of Translation) is an advanced neural framework developed to address the shortcomings of traditional machine translation (MT) evaluation metrics such as BLEU and METEOR. These conventional metrics primarily rely on surface-level lexical overlap between a machine-generated translation and a reference text, often missing deeper semantic nuances, especially when translations are highly fluent but non-literal. In contrast, COMET utilizes state-of-the-art pretrained cross-lingual language models, such as XLM-RoBERTa, enabling it to assess translation quality based on meaning rather than mere word matching.

One of COMET's defining features is its multilingual and cross-lingual adaptability. By employing multilingual pretrained models, it supports a wide range of languages, including those with limited available data. Furthermore, the framework is flexible enough to extend its capabilities to languages that were not directly represented during its training phase. This makes COMET particularly valuable for diverse linguistic contexts.

At its core, COMET is designed to emulate human judgment in evaluating translations. It is trained on human quality assessments, such as Direct Assessments (DA), where human evaluators score the adequacy and fluency of translations, and Human-mediated Translation Edit Rate (HTER), which measures the amount of effort required to correct a translation. By learning from these human-provided evaluations, COMET aligns closely with the way humans assess translation quality, allowing it to make reliable predictions at the segment level.

COMET is built around two primary model architectures: the Estimator model and the Translation Ranking

model . The Estimator model directly predicts a quality score by considering the machine translation output alongside a reference translation and, optionally, the source text. It is trained to regress on human judgment scores, such as DA or HTER, and provides a numeric score reflecting the quality of the translation. The Translation Ranking model, on the other hand, is designed to compare multiple translation hypotheses, ranking them based on their similarity to the reference and the source text. This model is particularly effective in differentiating between high-performing translations, where subtle differences in quality might not be captured by traditional metrics.

An additional strength of COMET lies in its flexibility to operate both in reference-based and reference-free modes. While traditional MT evaluation methods require a human-generated reference translation, COMET can function without one by leveraging the source text to generate a quality score. This reference-free evaluation capability is inspired by Quality Estimation (QE) models, which assess translation quality without needing a reference, further expanding COMET’s applicability in practical settings.

COMET’s technical implementation includes an innovative layer-wise attention mechanism , which pools information from various layers of its pretrained encoder model. Instead of relying solely on the final layer of the model, it assigns different weights to different layers, each capturing various linguistic features such as syntax and semantics. This comprehensive approach enhances COMET’s ability to make fine-grained predictions about translation quality.

In terms of performance, COMET has achieved state-of-the-art results . It was a standout in the WMT 2019 Metrics Shared Task , where it demonstrated a high level of correlation with human judgments, outperforming other metrics such as BLEU, CHRF, and BERTScore. Notably, COMET has proven to be particularly robust in evaluating high-performing MT systems, where traditional metrics often struggle to differentiate between similar outputs.

COMET is also widely used and available as an open-source tool , making it accessible to both researchers and professionals in the industry. Its flexibility allows for fine-tuning and customization, enabling it to be adapted for other natural language generation (NLG) tasks beyond machine translation, such as evaluating the coherence and factual accuracy of outputs from large language models.

In summary, COMET represents a significant advancement in the field of MT evaluation, offering a more nuanced, semantically driven approach to assessing translation quality. Its reliance on pretrained multilingual models, alignment with human judgment, and adaptability for reference-free evaluation make it a valuable tool in both research and practical applications.

### 4.3.3 Fine-tune NLI models

In the context of detecting hallucinated answers, we also employed Natural Language Inference (NLI) models, an approach that has seen substantial progress in recent years, particularly in its ability to capture and analyze semantic complexities that closely align with the phenomenon of hallucination. NLI models are designed to assess the logical relationship between two sentences, classifying the connection between a hypothesis and a premise into one of three categories: entailment, neutral, or contradiction. These categories map onto the model’s understanding of whether the hypothesis follows logically from the premise, is unrelated, or directly contradicts it.

Given their capacity to handle nuanced semantic relationships, NLI models offer a valuable framework for addressing the issue of hallucination detection in language models. Specifically, we framed the problem of hallucination detection as a task of NLI: given the input produced by the model, which we define as the hypothesis (denoted as *hyp*), and the corresponding target or reference (denoted as *tgt*), the task is to determine whether *tgt* entails, contradicts, or remains neutral with respect to *hyp*. In this context, hallucinated content would manifest itself as contradictions or neutral relationships, rather than entailments.

To implement this approach, we selected a pre-trained NLI model available through the Hugging Face library, specifically the mDeBERTa-v3-base architecture<sup>1</sup>, which has been trained on a large-scale multilingual dataset. The use of this model, which is based on DeBERTa’s third iteration, offers distinct advantages due to its capability to handle a broad range of linguistic structures and languages, making it highly adaptable to the varied and complex data we encountered.

---

<sup>1</sup><https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7>

To further tailor the NLI model to the specific demands of our task, we fine-tuned it using our annotated validation set. This step allowed us to calibrate the model to detect the subtle distinctions necessary for identifying hallucinations in generated outputs. By fine-tuning the model on a dataset that reflected the unique challenges of hallucination detection, we enhanced its ability to discern whether a model-generated output aligned with or deviated from the reference, effectively improving its performance in this specific domain.

This approach of converting hallucination detection into an NLI task not only provided a clear framework for assessing the logical coherence between the generated output and the reference but also allowed us to leverage the strengths of state-of-the-art NLI models. By applying these models, we were able to bring a structured, formal method to the task of hallucination detection, offering improved accuracy in identifying when generated text was consistent with the expected output or exhibited elements of hallucination.

#### 4.3.4 Fine-tune hallucination detection model

Our next technique involves the fine-tuning of a pre-trained classifier that is specifically designed for hallucination detection, with the goal of learning to differentiate between hallucinated and non-hallucinated instances within the SHROOM dataset. The primary objective of this approach is to enable the model to capture the subtle distinctions and patterns that characterize hallucinated content, thereby improving its ability to flag such occurrences in generated text.

To implement this, we utilized a pre-trained model based on the microsoft/deberta-v3-base architecture, provided through Hugging Face<sup>2</sup>. This particular model is uniquely suited for the task of hallucination detection due to its initial training on Natural Language Inference (NLI) data, which equips it with a strong foundation in determining textual entailment and relationships between sentences. In its pre-trained form, the model excels at identifying whether a hypothesis (generated text) is entailed, contradicted, or unrelated to a premise (reference text), which is a key element in identifying hallucinated content.

Following this initial training, the model underwent further fine-tuning using datasets focused on summarization tasks that include annotations related to factual consistency. These datasets are enriched with examples where the generated text is compared against reference summaries, with labels indicating whether the generated summaries maintain factual consistency or deviate into hallucination. This additional fine-tuning step is critical, as it helps the model adapt from a general NLI task to a more specialized domain of hallucination detection, particularly in the context of summarization and natural language generation tasks.

The output of this hallucination detection model is expressed as a probability score ranging from 0 to 1. A score closer to 0 indicates that the model has detected the presence of hallucinated content in the generated text, implying a significant divergence from factual accuracy. Conversely, a score approaching 1 suggests that the model perceives the generated content as factually consistent with the reference text or input data. The probabilistic nature of the model's output allows for a more nuanced evaluation of the content, providing a measure of the model's confidence in the factual accuracy of the generated hypotheses. This probabilistic framework is particularly valuable as it enables flexible thresholding, which can be adjusted depending on the desired level of tolerance for hallucinated content in different applications.

To tailor the model to the specific requirements of our task, we undertook a fine-tuning process using an annotated validation set comprising 1000 samples from the SHROOM dataset. This dataset had been pre-labeled with instances of hallucinated and non-hallucinated content, allowing the model to adapt to the specific intricacies of our domain. By fine-tuning the model on this domain-specific data, we aimed to improve its ability to detect hallucinations in SHROOM instances, accounting for the unique variability and complexity inherent in this dataset. This process enhanced the model's sensitivity to the patterns and characteristics typical of hallucinated content within the SHROOM domain, ensuring a more accurate detection of such instances.

Furthermore, we applied a thresholding approach to translate the model's probabilistic outputs into actionable decisions. Specifically, we set a threshold at 0.5, meaning that any prediction with a score greater than 0.5 is considered to indicate factual consistency between the input and output. On the other hand, predictions with a score below this threshold are flagged as potentially hallucinatory instances, signaling that the generated

<sup>2</sup>[https://huggingface.co/vectara/hallucination\\_evaluation\\_model](https://huggingface.co/vectara/hallucination_evaluation_model)

content may diverge from factual accuracy or truthfulness. This thresholding mechanism not only simplifies the interpretation of the model’s probabilistic scores but also provides a clear, binary classification for practical use.

By adopting this approach, we were able to systematically evaluate the output of the model and make consistent decisions regarding the presence of hallucinations. This method enhances the reliability and robustness of hallucination detection in the generated content, providing a structured and adaptable framework for improving the factual integrity of language model outputs.

### 4.3.5 Voting Classifier

In our final methodological approach, we implemented a sophisticated ensemble technique known as the Voting Classifier. This strategy is based on the concept of aggregating predictions from multiple models, combining their outputs to arrive at a more reliable and accurate result. The Voting Classifier works by gathering input from each of the individual classifiers—namely, the pre-trained hallucination detection model, its fine-tuned version, and the fine-tuned NLI model—and determining the final prediction based on a majority vote. This majority voting system ensures that the label assigned to each data point reflects the consensus of the ensemble, rather than relying on the potentially biased output of a single model.

The advantage of using a Voting Classifier lies in its ability to harness the unique strengths of each participating model, which may excel in different aspects of hallucination detection. For example, one model might be particularly adept at identifying syntactic inconsistencies, while another is better at detecting semantic misalignments. By combining these diverse perspectives, the Voting Classifier produces a more balanced and robust prediction, reducing the likelihood of error or misclassification. Additionally, this ensemble technique helps to counteract the weaknesses or limitations inherent in any one model. If one of the models performs poorly on a specific type of input, the other models can compensate, thereby enhancing the overall performance and making the system more resilient to the complexities of the dataset.

The ensemble approach is particularly beneficial in the context of hallucination detection, where the task involves recognizing subtle and varied types of discrepancies between the generated outputs and reference data. The inherent complexity and variability of such data make it challenging for a single model to consistently deliver high accuracy across all cases. By aggregating the predictions of multiple models, the Voting Classifier can better capture these nuanced differences, contributing to a more precise and comprehensive understanding of the hallucinations present in the dataset.

Moreover, the Voting Classifier not only improves binary classification decisions (i.e., determining whether a data point is hallucinated or not) but also adds flexibility through its probability-based outputs. In addition to the majority vote for binary labels, the classifier can calculate the percentage of models that flagged a given output as hallucinated, providing a clearer picture of the ensemble’s overall confidence in its predictions. This percentage-based approach gives users a more detailed insight into the degree of hallucination in a particular data point, making it easier to prioritize cases for further investigation or refinement.

By adopting this ensemble methodology, we have created a more resilient and adaptable system for hallucination detection. The Voting Classifier not only enhances predictive accuracy but also offers a more nuanced approach to understanding complex model-generated outputs, ensuring that our detection process is both reliable and scalable across a diverse range of data. This approach underscores our commitment to addressing the intricate challenges posed by hallucination detection and reinforces the robustness of our overall solution.

### 4.3.6 Prompting

In our final approach to detecting hallucinations, we implemented Large Language Model (LLM) Prompt Engineering as a sophisticated strategy to enhance the accuracy and reliability of generated outputs. This method leverages the advanced capabilities of state-of-the-art LLMs, enabling us to effectively generate, evaluate, and refine outputs systematically to identify and mitigate hallucinated content.

To achieve our objectives, we employed a diverse array of LLMs, including notable models such as Gemma and Llama. These models were selected for their distinct architectures and strengths, which provided us with a comprehensive toolkit for hallucination detection. Our methodology incorporated a variety of prompting

---

techniques designed to maximize the models’ potential for accurately identifying and addressing hallucinations. Specifically, we utilized zero-shot prompts, chain of thought reasoning, and few-shot learning strategies.

The zero-shot prompting technique enabled us to ask the models to perform tasks without prior examples, allowing them to rely on their trained knowledge to generate responses. This approach was particularly useful in scenarios where specific examples of hallucinations were scarce, allowing for broader generalization in the models’ outputs.

In addition, the chain of thought reasoning technique encouraged the models to engage in step-by-step reasoning, guiding them to break down complex problems into manageable parts. This method enhanced the models’ ability to track logical coherence and identify inconsistencies that may indicate hallucinations. By explicitly modeling the reasoning process, we fostered a deeper understanding of the relationships between concepts and improved the models’ critical evaluation of their outputs.

Furthermore, our implementation of few-shot learning strategies allowed us to provide the models with a limited number of examples to help them understand the context and nuances of the task. This adaptability enabled the models to learn from specific instances while minimizing the need for extensive supervision. By exposing the models to various examples of both accurate and hallucinated content, we empowered them to infer patterns and make informed judgments about the reliability of their generated responses.

Ultimately, our approach aims to harness the strengths of LLMs in understanding and generating human-like language while ensuring that the generated content remains factually accurate and free of hallucinations. By combining advanced prompting techniques with diverse LLMs, we are not only enhancing the models’ performance but also contributing to the broader discourse on responsible AI development and deployment. This work underscores the importance of rigorous evaluation methods in the field of natural language processing, paving the way for future advancements in ensuring the integrity and reliability of AI-generated content.



# Chapter 5

## Experiments

### Contents

---

|            |                                                |           |
|------------|------------------------------------------------|-----------|
| <b>5.1</b> | <b>Baseline System</b>                         | <b>72</b> |
| <b>5.2</b> | <b>Sentence Similarity</b>                     | <b>72</b> |
| 5.2.1      | Calculate Distance                             | 72        |
| 5.2.2      | Clustering                                     | 82        |
| <b>5.3</b> | <b>COMET</b>                                   | <b>84</b> |
| <b>5.4</b> | <b>NLI models</b>                              | <b>84</b> |
| <b>5.5</b> | <b>Fine-tune hallucination detection model</b> | <b>87</b> |
| <b>5.6</b> | <b>Voting Classifier</b>                       | <b>88</b> |
| <b>5.7</b> | <b>Prompting</b>                               | <b>90</b> |
| <b>5.8</b> | <b>Overall Results</b>                         | <b>94</b> |

---

## 5.1 Baseline System

Besides the dataset, the organizers of this shared task provided also a baseline system. As a baseline for the task, they used an LLM [22] to evaluate whether the generated hypotheses are coherent with the provided context. As described in the task description paper [35] the prompt template used derives from Self-checkgpt [32] as shown in 5.1.1. The system of [32], which has gathered some attention from the community, constitutes a straightforward approach based on a modern LLM, and is therefore well-suited to serve as a baseline in the shared-task: it corresponds to a reasonable default approach to tackle the problem the participants were challenged with.

```
Context: {}
Sentence: {}
Is the Sentence supported by the Context above?
Answer using ONLY yes or no:
```

Figure 5.1.1: Prompt template used in the baseline system

The specific context varies depending on the task addressed, i.e. the source sentence for the paraphrase generation task, and the target sentence for machine translation and definition modeling tasks. As for the probability of hallucination, the system relies on the probability assigned by the model to the first output word. In cases where the output does not clearly indicate *yes* or *no*, randomly one is selected, attributing a hallucination probability of 0.5.

The baseline system’s results are shown in Table 5.1 below.

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.65     | 0.38 |
| Val aware     | 0.71     | 0.46 |
| Test agnostic | 0.75     | 0.49 |
| Test aware    | 0.70     | 0.40 |

Table 5.1: Results of Baseline System

## 5.2 Sentence Similarity

Starting our experiments, we used various pre-trained transformer models to encode the ‘tgt’ and the ‘hyp’ of each datapoint in the validation dataset, both model-aware and model-agnostic datasets. The next step was to calculate the similarity score of these two sentences. to do so, we used a variety of distance calculating methods including Euclidean distance, manhattan distance, cosine similarity, and a few clustering methods.

### 5.2.1 Calculate Distance

In the initial phase of our experiments, we employed several pre-trained transformer models to encode both the target (‘tgt’) and the hypothesis (‘hyp’) for each data point in the validation dataset. This process was applied across both model-aware and model-agnostic datasets. The objective of this step was to compute a similarity score between the two sentences, representing the target and the hypothesis. To achieve this, we utilized various distance calculation methods, including Euclidean distance, Manhattan distance, cosine similarity, and a clustering algorithm.

For the computed distances, it was necessary to determine an appropriate threshold to perform binary classification, where each data point would be labeled as either a ‘Hallucination’ or ‘Not Hallucination.’ After conducting a series of exploratory experiments to identify a suitable threshold, we observed that higher classification accuracy was attained when the median of the distance values was used as the threshold. The median effectively separates the dataset into two halves, distinguishing the higher values from the lower ones. The thresholds derived for each model and dataset are presented in the App. A.

Once the thresholds were determined, the binary classification was carried out according to the following procedure:

- **For the Euclidean and Manhattan distances**, we initially normalized the computed distances to constrain their values within the range  $[0, 1]$ . Following normalization, a data point was classified as a 'Hallucination' if its normalized distance exceeded the threshold. In this case, the probability of hallucination, denoted as  $p(\text{'Hallucination'})$ , was set equal to the corresponding normalized distance.
- **For the cosine similarity**, which inherently yields a similarity score between 0 and 1, we compared the score against the threshold. If the similarity score was lower than the threshold, the data point was labeled as a 'Hallucination'. In this scenario, the probability of hallucination,  $p(\text{'Hallucination'})$ , was computed as one minus the similarity score (i.e.,  $p(\text{'Hallucination'}) = 1 - \text{similarity score}$ ).

The results for the accuracy metric are presented in the [5.2,5.3,5.4,5.5](#). Meanwhile, the results for the Spearman correlation metric are provided in [5.6,5.7,5.8,5.9](#)

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.61             | 0.59                  | 0.61             |
| stsb-roberta-large                                                   | 0.70             | 0.70                  | 0.70             |
| all-distilroberta-v1                                                 | 0.65             | 0.65                  | 0.65             |
| all-MiniLM-L6-v2                                                     | 0.66             | 0.66                  | 0.66             |
| all-MiniLM-L12-v2                                                    | 0.66             | 0.66                  | 0.66             |
| all-mpnet-base-v2                                                    | 0.67             | 0.67                  | 0.67             |
| all-roberta-large-v1                                                 | 0.65             | 0.65                  | 0.65             |
| paraphrase-MiniLM-L6-v2                                              | 0.66             | 0.67                  | 0.67             |
| paraphrase-mpnet-base-v2                                             | 0.70             | 0.70                  | 0.70             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.66             | 0.66                  | 0.66             |
| paraphrase-MiniLM-L12                                                | 0.67             | 0.67                  | 0.67             |
| paraphrase-MiniLM-L3                                                 | 0.65             | 0.66                  | 0.65             |
| paraphrase-TinyBERT-L6                                               | 0.68             | 0.68                  | 0.68             |
| paraphrase-distilroberta-base                                        | 0.66             | 0.69                  | 0.67             |
| xlm-r-distilroberta                                                  | 0.67             | 0.68                  | 0.67             |
| stsb-roberta-base-v2                                                 | 0.69             | 0.70                  | 0.69             |
| stsb-roberta-large                                                   | 0.70             | 0.70                  | 0.70             |
| stsb-distilroberta-base-v2                                           | 0.67             | 0.67                  | 0.68             |
| stsb-mpnet-base-v2                                                   | 0.70             | 0.70                  | 0.70             |
| multi-qa-distilbert-cos-v1                                           | 0.66             | 0.66                  | 0.66             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.66             | 0.66                  | 0.66             |
| sentence-t5-base                                                     | 0.71             | 0.71                  | 0.71             |
| sentence-t5-large                                                    | <b>0.72</b>      | <b>0.72</b>           | <b>0.72</b>      |
| gtr-t5-base                                                          | 0.68             | 0.68                  | 0.68             |
| gtr-t5-large                                                         | 0.68             | 0.68                  | 0.68             |
| nq-distilbert-base-v1                                                | 0.65             | 0.65                  | 0.65             |
| nli-distilroberta-base-v2                                            | 0.68             | 0.69                  | 0.68             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.68             | 0.67                  | 0.67             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.69             | 0.7                   | 0.69             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.63             | 0.63                  | 0.64             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.67             | 0.68                  | 0.67             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.68             | 0.67                  | 0.67             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.68             | 0.67                  | 0.67             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.68             | 0.68                  | 0.67             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.63             | 0.63                  | 0.64             |

Table 5.2: Accuracy of model-agnostic test set

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.55             | 0.55                  | 0.54             |
| stsb-roberta-large                                                   | 0.59             | 0.6                   | 0.6              |
| all-distilroberta-v1                                                 | 0.58             | 0.58                  | 0.59             |
| all-MiniLM-L6-v2                                                     | 0.57             | 0.57                  | 0.57             |
| all-MiniLM-L12-v2                                                    | 0.58             | 0.58                  | 0.58             |
| all-mpnet-base-v2                                                    | 0.58             | 0.58                  | 0.58             |
| all-roberta-large-v1                                                 | 0.59             | 0.59                  | 0.58             |
| paraphrase-MiniLM-L6-v2                                              | 0.56             | 0.58                  | 0.56             |
| paraphrase-mpnet-base-v2                                             | 0.59             | 0.59                  | 0.59             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.57             | 0.57                  | 0.56             |
| paraphrase-MiniLM-L12                                                | 0.56             | 0.57                  | 0.56             |
| paraphrase-MiniLM-L3                                                 | 0.55             | 0.57                  | 0.56             |
| paraphrase-TinyBERT-L6                                               | 0.56             | 0.58                  | 0.56             |
| paraphrase-distilroberta-base                                        | 0.57             | 0.59                  | 0.57             |
| xlm-r-distilroberta                                                  | 0.59             | 0.58                  | 0.58             |
| stsb-roberta-base-v2                                                 | 0.57             | 0.58                  | 0.57             |
| stsb-roberta-large                                                   | 0.59             | 0.60                  | 0.60             |
| stsb-distilroberta-base-v2                                           | 0.57             | 0.57                  | 0.57             |
| stsb-mpnet-base-v2                                                   | 0.58             | 0.59                  | 0.58             |
| multi-qa-distilbert-cos-v1                                           | 0.57             | 0.57                  | 0.57             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.57             | 0.57                  | 0.56             |
| sentence-t5-base                                                     | 0.60             | 0.60                  | 0.60             |
| sentence-t5-large                                                    | <b>0.62</b>      | <b>0.62</b>           | <b>0.62</b>      |
| gtr-t5-base                                                          | 0.58             | 0.58                  | 0.58             |
| gtr-t5-large                                                         | 0.59             | 0.59                  | 0.59             |
| nq-distilbert-base-v1                                                | 0.58             | 0.58                  | 0.58             |
| nli-distilroberta-base-v2                                            | 0.58             | 0.58                  | 0.58             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.57             | 0.58                  | 0.57             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.58             | 0.58                  | 0.58             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.56             | 0.56                  | 0.56             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.59             | 0.58                  | 0.58             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.57             | 0.57                  | 0.57             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.57             | 0.57                  | 0.57             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.58             | 0.59                  | 0.58             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.56             | 0.56                  | 0.56             |

Table 5.3: Accuracy of model-aware test set

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.67             | 0.65                  | 0.67             |
| stsb-roberta-large                                                   | 0.74             | 0.74                  | 0.74             |
| all-distilroberta-v1                                                 | 0.68             | 0.68                  | 0.68             |
| all-MiniLM-L6-v2                                                     | 0.69             | 0.69                  | 0.69             |
| all-MiniLM-L12-v2                                                    | 0.69             | 0.69                  | 0.69             |
| all-mpnet-base-v2                                                    | 0.68             | 0.68                  | 0.68             |
| all-roberta-large-v1                                                 | 0.71             | 0.71                  | 0.71             |
| paraphrase-MiniLM-L6-v2                                              | 0.74             | 0.71                  | 0.74             |
| paraphrase-mpnet-base-v2                                             | 0.72             | 0.72                  | 0.72             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.71             | 0.71                  | 0.71             |
| paraphrase-MiniLM-L12                                                | 0.74             | 0.72                  | 0.74             |
| paraphrase-MiniLM-L3                                                 | 0.70             | 0.70                  | 0.69             |
| paraphrase-TinyBERT-L6                                               | 0.71             | 0.72                  | 0.7              |
| paraphrase-distilroberta-base                                        | 0.72             | 0.73                  | 0.72             |
| xlm-r-distilroberta                                                  | 0.72             | 0.73                  | 0.72             |
| stsb-roberta-base-v2                                                 | 0.75             | 0.75                  | 0.75             |
| stsb-roberta-large                                                   | 0.74             | 0.74                  | 0.74             |
| stsb-distilroberta-base-v2                                           | 0.73             | 0.74                  | 0.72             |
| stsb-mpnet-base-v2                                                   | <b>0.75</b>      | <b>0.75</b>           | <b>0.75</b>      |
| multi-qa-distilbert-cos-v1                                           | 0.69             | 0.69                  | 0.69             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.71             | 0.71                  | 0.71             |
| sentence-t5-base                                                     | 0.73             | 0.73                  | 0.73             |
| sentence-t5-large                                                    | 0.74             | 0.74                  | 0.74             |
| gtr-t5-base                                                          | 0.69             | 0.69                  | 0.69             |
| gtr-t5-large                                                         | 0.70             | 0.70                  | 0.70             |
| nq-distilbert-base-v1                                                | 0.69             | 0.69                  | 0.69             |
| nli-distilroberta-base-v2                                            | 0.72             | 0.72                  | 0.72             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.73             | 0.72                  | 0.73             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.73             | 0.73                  | 0.74             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.66             | 0.66                  | 0.67             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.72             | 0.73                  | 0.72             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.72             | 0.72                  | 0.72             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.72             | 0.72                  | 0.72             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.72             | 0.73                  | 0.73             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.66             | 0.66                  | 0.67             |

Table 5.4: Accuracy of model-agnostic validation set

| Model                                                                | Acc<br>Euclidean | Acc Cos<br>Similarity | Acc<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.53             | 0.53                  | 0.53             |
| stsb-roberta-large                                                   | 0.58             | 0.59                  | 0.58             |
| all-distilroberta-v1                                                 | 0.57             | 0.57                  | 0.56             |
| all-MiniLM-L6-v2                                                     | 0.56             | 0.56                  | 0.56             |
| all-MiniLM-L12-v2                                                    | 0.57             | 0.57                  | 0.57             |
| all-mpnet-base-v2                                                    | 0.56             | 0.56                  | 0.56             |
| all-roberta-large-v1                                                 | 0.57             | 0.57                  | 0.57             |
| paraphrase-MiniLM-L6-v2                                              | 0.55             | 0.57                  | 0.55             |
| paraphrase-mpnet-base-v2                                             | 0.57             | 0.59                  | 0.57             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.55             | 0.55                  | 0.55             |
| paraphrase-MiniLM-L12                                                | 0.57             | 0.57                  | 0.57             |
| paraphrase-MiniLM-L3                                                 | 0.54             | 0.56                  | 0.54             |
| paraphrase-TinyBERT-L6                                               | 0.55             | 0.56                  | 0.55             |
| paraphrase-distilroberta-base                                        | 0.55             | 0.55                  | 0.55             |
| xlm-r-distilroberta                                                  | 0.57             | 0.57                  | 0.56             |
| stsb-roberta-base-v2                                                 | 0.56             | 0.58                  | 0.56             |
| stsb-roberta-large                                                   | 0.58             | 0.59                  | 0.58             |
| stsb-distilroberta-base-v2                                           | 0.55             | 0.56                  | 0.56             |
| stsb-mpnet-base-v2                                                   | 0.56             | 0.56                  | 0.57             |
| multi-qa-distilbert-cos-v1                                           | 0.55             | 0.55                  | 0.55             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.55             | 0.55                  | 0.55             |
| sentence-t5-base                                                     | 0.58             | 0.58                  | 0.58             |
| sentence-t5-large                                                    | <b>0.61</b>      | <b>0.61</b>           | <b>0.6</b>       |
| gtr-t5-base                                                          | 0.56             | 0.56                  | 0.55             |
| gtr-t5-large                                                         | 0.57             | 0.57                  | 0.57             |
| nq-distilbert-base-v1                                                | 0.56             | 0.57                  | 0.56             |
| nli-distilroberta-base-v2                                            | 0.57             | 0.57                  | 0.57             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.55             | 0.58                  | 0.55             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.58             | 0.59                  | 0.57             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.53             | 0.53                  | 0.54             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.57             | 0.57                  | 0.56             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.55             | 0.55                  | 0.55             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.55             | 0.55                  | 0.55             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.57             | 0.58                  | 0.57             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.53             | 0.53                  | 0.54             |

Table 5.5: Accuracy of model-aware validation set

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.35             | 0.33                  | 0.36             |
| stsb-roberta-large                                                   | 0.56             | 0.56                  | 0.56             |
| all-distilroberta-v1                                                 | 0.43             | 0.43                  | 0.43             |
| all-MiniLM-L6-v2                                                     | 0.43             | 0.43                  | 0.43             |
| all-MiniLM-L12-v2                                                    | 0.44             | 0.44                  | 0.44             |
| all-mpnet-base-v2                                                    | 0.47             | 0.47                  | 0.47             |
| all-roberta-large-v1                                                 | 0.45             | 0.45                  | 0.45             |
| paraphrase-MiniLM-L6-v2                                              | 0.46             | 0.47                  | 0.46             |
| paraphrase-mpnet-base-v2                                             | 0.54             | 0.53                  | 0.54             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.46             | 0.46                  | 0.46             |
| paraphrase-MiniLM-L12                                                | 0.5              | 0.49                  | 0.51             |
| paraphrase-MiniLM-L3                                                 | 0.41             | 0.43                  | 0.41             |
| paraphrase-TinyBERT-L6                                               | 0.48             | 0.5                   | 0.48             |
| paraphrase-distilroberta-base                                        | 0.48             | 0.5                   | 0.48             |
| xlm-r-distilroberta                                                  | 0.48             | 0.49                  | 0.48             |
| stsb-roberta-base-v2                                                 | 0.54             | 0.56                  | 0.54             |
| stsb-roberta-large                                                   | 0.56             | 0.56                  | 0.56             |
| stsb-distilroberta-base-v2                                           | 0.49             | 0.51                  | 0.49             |
| stsb-mpnet-base-v2                                                   | 0.56             | 0.56                  | 0.56             |
| multi-qa-distilbert-cos-v1                                           | 0.46             | 0.46                  | 0.46             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.46             | 0.46                  | 0.46             |
| sentence-t5-base                                                     | 0.57             | 0.57                  | 0.57             |
| sentence-t5-large                                                    | 0.59             | 0.59                  | 0.59             |
| gtr-t5-base                                                          | 0.48             | 0.48                  | 0.48             |
| gtr-t5-large                                                         | 0.5              | 0.5                   | 0.5              |
| nq-distilbert-base-v1                                                | 0.42             | 0.42                  | 0.42             |
| nli-distilroberta-base-v2                                            | 0.49             | 0.52                  | 0.49             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.5              | 0.49                  | 0.5              |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.54             | 0.55                  | 0.54             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.41             | 0.41                  | 0.41             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.48             | 0.49                  | 0.48             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.49             | 0.49                  | 0.49             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.49             | 0.49                  | 0.49             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.5              | 0.51                  | 0.5              |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.41             | 0.41                  | 0.41             |

Table 5.6: Spearman Correlation of model-agnostic test set

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.04             | 0.04                  | 0.05             |
| stsb-roberta-large                                                   | 0.11             | 0.12                  | 0.2              |
| all-distilroberta-v1                                                 | 0.12             | 0.12                  | 0.14             |
| all-MiniLM-L6-v2                                                     | 0.1              | 0.1                   | 0.11             |
| all-MiniLM-L12-v2                                                    | 0.11             | 0.11                  | 0.14             |
| all-mpnet-base-v2                                                    | 0.1              | 0.1                   | 0.1              |
| all-roberta-large-v1                                                 | 0.12             | 0.12                  | 0.12             |
| paraphrase-MiniLM-L6-v2                                              | 0.12             | 0.11                  | 0.16             |
| paraphrase-mpnet-base-v2                                             | 0.11             | 0.12                  | 0.11             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.12             |
| paraphrase-MiniLM-L12                                                | 0.09             | 0.11                  | 0.11             |
| paraphrase-MiniLM-L3                                                 | 0.06             | 0.11                  | 0.07             |
| paraphrase-TinyBERT-L6                                               | 0.14             | 0.12                  | 0.16             |
| paraphrase-distilroberta-base                                        | 0.13             | 0.13                  | 0.14             |
| xlm-r-distilroberta                                                  | 0.1              | 0.1                   | 0.1              |
| stsb-roberta-base-v2                                                 | 0.16             | 0.13                  | 0.2              |
| stsb-roberta-large                                                   | 0.11             | 0.12                  | 0.2              |
| stsb-distilroberta-base-v2                                           | 0.11             | 0.11                  | 0.16             |
| stsb-mpnet-base-v2                                                   | 0.11             | 0.14                  | 0.15             |
| multi-qa-distilbert-cos-v1                                           | 0.1              | 0.1                   | 0.11             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.12             |
| sentence-t5-base                                                     | 0.15             | 0.15                  | 0.15             |
| sentence-t5-large                                                    | 0.16             | 0.16                  | 0.16             |
| gtr-t5-base                                                          | 0.15             | 0.15                  | 0.15             |
| gtr-t5-large                                                         | 0.14             | 0.14                  | 0.14             |
| nq-distilbert-base-v1                                                | 0.15             | 0.11                  | 0.18             |
| nli-distilroberta-base-v2                                            | 0.1              | 0.11                  | 0.13             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.15             | 0.11                  | 0.2              |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.12             | 0.13                  | 0.12             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.12             | 0.06                  | 0.13             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.1              | 0.1                   | 0.1              |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.14             | 0.13                  | 0.14             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.14             | 0.13                  | 0.14             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.15             | 0.14                  | 0.16             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.12             | 0.06                  | 0.13             |

Table 5.7: Spearman Correlation of model-aware test set

| <b>Model</b>                                                         | <b>Rho<br/>Euclidean</b> | <b>Rho Cos<br/>Similarity</b> | <b>Rho<br/>Manhattan</b> |
|----------------------------------------------------------------------|--------------------------|-------------------------------|--------------------------|
| bert-base-uncased                                                    | 0.44                     | 0.43                          | 0.44                     |
| stsb-roberta-large                                                   | 0.63                     | 0.63                          | 0.63                     |
| all-distilroberta-v1                                                 | 0.53                     | 0.53                          | 0.52                     |
| all-MiniLM-L6-v2                                                     | 0.54                     | 0.54                          | 0.54                     |
| all-MiniLM-L12-v2                                                    | 0.53                     | 0.53                          | 0.53                     |
| all-mpnet-base-v2                                                    | 0.55                     | 0.55                          | 0.54                     |
| all-roberta-large-v1                                                 | 0.54                     | 0.54                          | 0.54                     |
| paraphrase-MiniLM-L6-v2                                              | 0.57                     | 0.57                          | 0.57                     |
| paraphrase-mpnet-base-v2                                             | 0.62                     | 0.62                          | 0.61                     |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.56                     | 0.56                          | 0.56                     |
| paraphrase-MiniLM-L12                                                | 0.61                     | 0.59                          | 0.61                     |
| paraphrase-MiniLM-L3                                                 | 0.52                     | 0.53                          | 0.52                     |
| paraphrase-TinyBERT-L6                                               | 0.56                     | 0.59                          | 0.56                     |
| paraphrase-distilroberta-base                                        | 0.6                      | 0.62                          | 0.6                      |
| xlm-r-distilroberta                                                  | 0.57                     | 0.59                          | 0.57                     |
| stsb-roberta-base-v2                                                 | 0.63                     | 0.63                          | 0.63                     |
| stsb-roberta-large                                                   | 0.63                     | 0.63                          | 0.63                     |
| stsb-distilroberta-base-v2                                           | 0.6                      | 0.62                          | 0.6                      |
| stsb-mpnet-base-v2                                                   | 0.64                     | 0.64                          | 0.64                     |
| multi-qa-distilbert-cos-v1                                           | 0.56                     | 0.56                          | 0.56                     |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.56                     | 0.56                          | 0.56                     |
| sentence-t5-base                                                     | 0.63                     | 0.63                          | 0.63                     |
| sentence-t5-large                                                    | 0.65                     | 0.65                          | 0.65                     |
| gtr-t5-base                                                          | 0.56                     | 0.56                          | 0.56                     |
| gtr-t5-large                                                         | 0.58                     | 0.58                          | 0.58                     |
| nq-distilbert-base-v1                                                | 0.53                     | 0.53                          | 0.53                     |
| nli-distilroberta-base-v2                                            | 0.6                      | 0.63                          | 0.6                      |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.62                     | 0.6                           | 0.62                     |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.64                     | 0.63                          | 0.63                     |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.54                     | 0.53                          | 0.54                     |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.57                     | 0.59                          | 0.57                     |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.58                     | 0.58                          | 0.58                     |
| sentence-transformers-quora-distilbert-multilingual                  | 0.58                     | 0.58                          | 0.58                     |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.62                     | 0.62                          | 0.62                     |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.54                     | 0.53                          | 0.54                     |

Table 5.8: Spearman Correlation of model-agnostic validation set

| Model                                                                | Rho<br>Euclidean | Rho Cos<br>Similarity | Rho<br>Manhattan |
|----------------------------------------------------------------------|------------------|-----------------------|------------------|
| bert-base-uncased                                                    | 0.05             | 0.06                  | 0.07             |
| stsb-roberta-large                                                   | 0.13             | 0.13                  | 0.23             |
| all-distilroberta-v1                                                 | 0.12             | 0.12                  | 0.15             |
| all-MiniLM-L6-v2                                                     | 0.1              | 0.1                   | 0.12             |
| all-MiniLM-L12-v2                                                    | 0.11             | 0.11                  | 0.16             |
| all-mpnet-base-v2                                                    | 0.11             | 0.11                  | 0.11             |
| all-roberta-large-v1                                                 | 0.12             | 0.12                  | 0.12             |
| paraphrase-MiniLM-L6-v2                                              | 0.14             | 0.13                  | 0.18             |
| paraphrase-mpnet-base-v2                                             | 0.13             | 0.14                  | 0.13             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.13             |
| paraphrase-MiniLM-L12                                                | 0.1              | 0.12                  | 0.13             |
| paraphrase-MiniLM-L3                                                 | 0.07             | 0.12                  | 0.07             |
| paraphrase-TinyBERT-L6                                               | 0.15             | 0.13                  | 0.17             |
| paraphrase-distilroberta-base                                        | 0.15             | 0.16                  | 0.16             |
| xlm-r-distilroberta                                                  | 0.11             | 0.12                  | 0.11             |
| stsb-roberta-base-v2                                                 | 0.17             | 0.14                  | 0.21             |
| stsb-roberta-large                                                   | 0.13             | 0.13                  | 0.23             |
| stsb-distilroberta-base-v2                                           | 0.16             | 0.15                  | 0.2              |
| stsb-mpnet-base-v2                                                   | 0.13             | 0.15                  | 0.15             |
| multi-qa-distilbert-cos-v1                                           | 0.1              | 0.1                   | 0.11             |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.1              | 0.1                   | 0.13             |
| sentence-t5-base                                                     | 0.17             | 0.17                  | 0.17             |
| sentence-t5-large                                                    | 0.19             | 0.19                  | 0.19             |
| gtr-t5-base                                                          | 0.16             | 0.16                  | 0.16             |
| gtr-t5-large                                                         | 0.15             | 0.15                  | 0.16             |
| nq-distilbert-base-v1                                                | 0.17             | 0.12                  | 0.19             |
| nli-distilroberta-base-v2                                            | 0.13             | 0.15                  | 0.17             |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.16             | 0.13                  | 0.21             |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.14             | 0.15                  | 0.14             |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.15             | 0.08                  | 0.16             |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.11             | 0.12                  | 0.11             |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.15             | 0.14                  | 0.15             |
| sentence-transformers-quora-distilbert-multilingual                  | 0.15             | 0.14                  | 0.15             |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.16             | 0.16                  | 0.17             |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.15             | 0.08                  | 0.16             |

Table 5.9: Spearman Correlation of model-aware validation set

### 5.2.2 Clustering

Another method we employed to enhance our classification process was clustering. Specifically, we calculated the distance between the target (tgt) and hypothesis (hyp) by determining the absolute difference between their respective vector representations. This metric served as a foundation for clustering the data points. To this end, we utilized the K-means algorithm, which is well-suited for partitioning data into distinct clusters based on similarity. By setting the number of clusters to two, we aimed to categorize the target-hypothesis pairs into two groups, thereby identifying patterns or commonalities among the data.

The clustering results were consistent with the findings obtained through the distance-based classification method described in the previous subsection. As demonstrated in Table 5.10, the clustering approach provided similar categorizations, further reinforcing the robustness of the distance calculation method. These comparable results suggest that both techniques are valid and complementary for this task, offering different perspectives on the underlying data distribution while yielding congruent classifications.

| Model                                                                | Test-agnostic | Test-aware | Val-agnostic | Val-aware |
|----------------------------------------------------------------------|---------------|------------|--------------|-----------|
| bert-base-uncased                                                    | 0.61          | 0.52       | 0.67         | 0.56      |
| stsb-roberta-large                                                   | 0.7           | 0.53       | 0.73         | 0.56      |
| all-distilroberta-v1                                                 | 0.65          | 0.52       | 0.65         | 0.55      |
| all-MiniLM-L6-v2                                                     | 0.64          | 0.52       | 0.69         | 0.54      |
| all-MiniLM-L12-v2                                                    | 0.66          | 0.53       | 0.67         | 0.56      |
| all-mpnet-base-v2                                                    | 0.66          | 0.51       | 0.68         | 0.52      |
| all-roberta-large-v1                                                 | 0.65          | 0.57       | 0.7          | 0.55      |
| paraphrase-MiniLM-L6-v2                                              | 0.66          | 0.52       | 0.73         | 0.54      |
| paraphrase-mpnet-base-v2                                             | 0.69          | 0.52       | 0.71         | 0.56      |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.65          | 0.53       | 0.7          | 0.55      |
| paraphrase-MiniLM-L12                                                | 0.67          | 0.53       | 0.73         | 0.56      |
| paraphrase-MiniLM-L3                                                 | 0.65          | 0.53       | 0.69         | 0.56      |
| paraphrase-TinyBERT-L6                                               | 0.67          | 0.56       | 0.70         | 0.54      |
| paraphrase-distilroberta-base                                        | 0.66          | 0.56       | 0.72         | 0.54      |
| xlm-r-distilroberta                                                  | 0.68          | 0.52       | 0.72         | 0.566     |
| stsb-roberta-base-v2                                                 | 0.7           | 0.57       | 0.74         | 0.56      |
| stsb-roberta-large                                                   | 0.7           | 0.53       | 0.73         | 0.56      |
| stsb-distilroberta-base-v2                                           | 0.67          | 0.51       | 0.72         | 0.54      |
| stsb-mpnet-base-v2                                                   | 0.70          | 0.53       | 0.75         | 0.56      |
| multi-qa-distilbert-cos-v1                                           | 0.64          | 0.52       | 0.68         | 0.53      |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.65          | 0.53       | 0.70         | 0.55      |
| sentence-t5-base                                                     | 0.7           | 0.58       | 0.73         | 0.57      |
| sentence-t5-large                                                    | 0.71          | 0.59       | 0.73         | 0.59      |
| gtr-t5-base                                                          | 0.64          | 0.57       | 0.68         | 0.55      |
| gtr-t5-large                                                         | 0.67          | 0.59e      | 0.7          | 0.57      |
| nq-distilbert-base-v1                                                | 0.64          | 0.52       | 0.68         | 0.54      |
| nli-distilroberta-base-v2                                            | 0.67          | 0.52       | 0.73         | 0.55      |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.68          | 0.52       | 0.73         | 0.55      |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.69          | 0.54       | 0.73         | 0.53      |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.63          | 0.55       | 0.66         | 0.53      |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.68          | 0.52       | 0.72         | 0.56      |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.68          | 0.56       | 0.72         | 0.54      |
| sentence-transformers-quora-distilbert-multilingual                  | 0.68          | 0.56       | 0.72         | 0.54      |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.68          | 0.57       | 0.71         | 0.53      |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.63          | 0.55       | 0.66         | 0.53      |

Table 5.10: Kmeans Accuracy

### 5.3 COMET

As previously discussed, COMET focuses on Machine Translate task but it can be used for different NLG tasks. So, our focus was exclusively directed towards the task of translation. To evaluate the quality of the generated translations, we employed COMET, a widely recognized and robust benchmark in the field of machine translation. COMET assesses translation quality by leveraging both the reference translation and the source, offering a comprehensive performance measure.

To conduct a detailed evaluation, we first distinguished the relevant data points specific to the translation task. Using these distinct data points, we applied the COMET benchmark to derive a quantitative score, which ranges from zero to one, with higher values indicating a superior translation quality. This evaluation method provided an objective measure of how well the models performed in terms of accuracy, fluency, and overall translation quality. The results of this evaluation, as obtained from our experiments, are presented in Table 1.12, offering insights into the comparative performance of the models under examination.

| Dataset       | Accuracy | Rho   |
|---------------|----------|-------|
| Val agnostic  | 0.82     | 0.718 |
| Val aware     | 0.71     | 0.64  |
| Test agnostic | 0.78     | 0.6   |
| Test aware    | 0.75     | 0.61  |

Table 5.11: Accuracy and Spearman Correlation on MT task using COMET, with a threshold of 0.75

Continuing our experiments we tested COMET in the rest tasks (DM and PG) and the results are shown in Tables 5.12 and 5.13

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.68     | 0.48 |
| Val aware     | 0.68     | 0.54 |
| Test agnostic | 0.69     | 0.5  |
| Test aware    | 0.65     | 0.46 |

Table 5.12: Accuracy and Spearman Correlation on DM task using COMET, with a threshold of 0.55

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.68     | 0.54 |
| Val aware     | 0.8      | 0.27 |
| Test agnostic | 0.74     | 0.47 |
| Test aware    | 0.84     | 0.42 |

Table 5.13: Accuracy and Spearman Correlation on PG task using COMET, with a threshold of 0.55

After calculating the results for each task separately, we combined them to get a final score for the whole dataset. These results are shown in Table 5.14

### 5.4 NLI models

In this section, we outline the experiments conducted to evaluate the effectiveness of fine-tuned Natural Language Inference (NLI) models in detecting hallucinations produced by large language models (LLMs). We employed pre-trained NLI models and further fine-tuned them using a task-specific dataset curated to better capture hallucination instances. By aligning these models to the task of hallucination detection, we

| Dataset       | Accuracy | Rho  |
|---------------|----------|------|
| Val agnostic  | 0.74     | 0.55 |
| Val aware     | 0.72     | 0.53 |
| Test agnostic | 0.73     | 0.53 |
| Test aware    | 0.74     | 0.53 |

Table 5.14: Accuracy and Spearman Correlation on all tasks using COMET, with different threshold per task

trained them to assess the relationship between a target sentence (ground truth) and a hypothesis (output from the LLM). The hypothesis is then classified as either entailed, contradicted, or neutral concerning the target, providing a framework to detect inconsistencies or fabricated content.

We utilized a pre-trained NLI model from the Hugging Face library, specifically the mDeBERTa-v3-base architecture. This model, part of the third iteration of the DeBERTa (Decoding-enhanced BERT with disentangled attention) family, was trained on a large multilingual dataset, making it an excellent candidate for our work. The pre-training on such a large-scale dataset enabled the model to capture a broad spectrum of linguistic features and patterns across various languages, offering a solid foundation for processing a wide array of linguistic structures. This was crucial for handling the complexity and variability of the data we worked with, both in terms of linguistic diversity and content characteristics.

Thanks to its disentangled attention mechanism, the mDeBERTa-v3-base model’s ability to manage multilingual inputs effectively made it highly adaptable to tasks involving different languages and dialects. This mechanism allows the model to better separate positional information from content, enhancing its understanding of language, and making it particularly suited to our task of processing intricate and diverse linguistic data.

The NLI model had already been trained on the multilingual-nli-26lang-2mil7 dataset and the XNLI validation dataset, both containing three labels: ‘entailment,’ ‘neutral,’ and ‘contradiction.’ During training, we mapped the ‘Hallucination’ label to ‘contradiction’ and the ‘Not Hallucination’ label to ‘entailment,’ creating a binary classification to clearly represent the hallucinatory nature of the content. This transformation simplified the learning process by providing distinct labels for the model to differentiate between hallucinatory and non-hallucinatory instances.

However, we didn’t solely rely on the pre-trained model’s capabilities. To address the specific demands of our project, we fine-tuned the model using a carefully annotated validation set that reflected the unique requirements of hallucination detection. This fine-tuning process enabled the model to detect subtle distinctions necessary for identifying hallucinations in model-generated outputs. In this context, hallucinations refer to cases where a model produces content that is not grounded in or supported by the original reference data—a common issue in text generation tasks.

By fine-tuning the mDeBERTa-v3-base architecture on a dataset designed specifically to address these challenges, we enhanced its sensitivity to both syntactic and semantic discrepancies. This additional training allowed the model to more effectively determine whether a generated output was aligned with the reference text or if it deviated in subtle, often misleading ways. This improvement not only boosted the model’s overall performance in detecting hallucinations but also increased its reliability in handling the specific types of data we encountered, resulting in significant improvements in accuracy and reliability.

Ultimately, combining a powerful pre-trained model with task-specific fine-tuning enabled us to develop a highly specialized tool for detecting hallucinations in generated outputs. This careful calibration enhanced the model’s ability to operate effectively within the context of our work, ensuring it could make the precise distinctions needed to improve the overall quality and trustworthiness of the generated content.

Following training, the model’s predictions were assessed using an entailment score, and a carefully chosen threshold was applied to distinguish between hallucinations and non-hallucinations. Through experimentation, we determined that a threshold of 0.8 optimized the accuracy of our trial set. Additionally, for

calculating the percentage of hallucinations for each data point, we subtracted the entailment percentage from 1.

The hyperparameters utilized for the NLI model fine-tuning mirrored those employed during the training of the initial model. The selection of hyperparameters followed a series of experiments, which yielded significantly lower levels of accuracy. Some of the experiments are displayed in Table 5.15 while a detailed account of the parameters employed for training this NLI model is outlined in Table 5.16.

| epochs | lr    | warmup ratio | weight decay | accuracy |
|--------|-------|--------------|--------------|----------|
| 5      | 2e-05 | 0.06         | 0.01         | 0.83     |
| 10     | 2e-06 | 0.1          | 0.01         | 0.75     |
| 5      | 2e-04 | 0.01         | 0.05         | 0.53     |
| 5      | 2e-05 | 0.05         | 0.001        | 0.8      |
| 5      | 2e-06 | 0.08         | 0.1          | 0.79     |

Table 5.15: Accuracy on trial-set from experiments with hyperparameters. The first row displays the hyperparameters chosen for finetuning

| Hyperparameter | Value                          |
|----------------|--------------------------------|
| train dataset  | validation set (1,000 samples) |
| learning rate  | 2e-05                          |
| epochs         | 5                              |
| warm-up ratio  | 0.06                           |
| weight decay   | 0.01                           |

Table 5.16: Hyperparameters used for NLI fine-tuning

The results before fine-tuning and after are shown in the Table 5.17

|                   | test agn | test aware | val agn | val aware | trial |
|-------------------|----------|------------|---------|-----------|-------|
| Before finetuning | 0.744    | 0.736      | 0.772   | 0.735     | 0.812 |
| After fine-tuning | 0.751    | 0.77       | -       | -         | 0.83  |

Table 5.17: Accuracy of nli models before and after finetuning

## 5.5 Fine-tune hallucination detection model

Our fine-tuned model underwent a comprehensive training and evaluation process, leveraging the SHROOM data provided by the task organizers. This data was essential in training the model to detect hallucinations in generated outputs accurately. Specifically, the model was trained using an annotated validation set, which contained ground-truth labels for hallucination detection, and its performance was then assessed using a trial set to ensure it generalized well to unseen data.

In the pre-processing phase, each data point was carefully prepared for input into the model. We extracted two key components from each instance: the hypothesis (*hyp*) and the target (*tgt*). The *hyp* represented the generated output from the model that needed to be evaluated for hallucinations, while the *tgt* served as the reference or ground truth, providing the basis for comparison. These components were then fed into the model to allow it to assess whether the generated output aligned with the reference text or deviated in ways that indicated hallucination.

To further enhance the model’s performance in terms of both accuracy and its ability to predict the probability of hallucination ( $p(\text{‘Hallucination’})$ ), we implemented a dual-training strategy. This strategy was crucial for optimizing the model’s predictions and improving its flexibility. In the first iteration, we trained the model using binary labels: ‘0’ for hallucination and ‘1’ for non-hallucination. This binary classification allowed the model to make straightforward decisions about whether or not a given hypothesis contained hallucinated content.

In the second iteration, we took a more nuanced approach by using float labels that represented the complement of the probability of hallucination, or  $(1-p(\text{‘Hallucination’}))$ . This allowed the model to produce a more detailed probability score, reflecting the likelihood of hallucination in the generated output rather than simply assigning a binary label. By employing these continuous values, the model could better capture the degree of uncertainty or confidence associated with its predictions.

This dual-training approach offered several advantages. First, it enabled the model to simultaneously provide a binary label indicating the presence or absence of hallucination and a probability score that quantified the likelihood of hallucination. This allowed us to balance between making categorical decisions about hallucinations while also gaining a deeper insight into the model’s confidence in its predictions. The probability score was particularly useful for scenarios where a more granular understanding of the model’s output was needed, such as when further human review or refinement was required.

Additionally, this approach provided flexibility in model deployment, as we could tailor the system to different use cases. In some instances, the binary label would suffice, especially in high-precision applications where a clear-cut decision is essential. In others, the probability score could be leveraged to prioritize or flag outputs for closer inspection, especially when the model indicated a moderate likelihood of hallucination.

Overall, this combination of binary and probabilistic outputs, achieved through the dual-training process, enhanced the model’s adaptability and effectiveness in detecting hallucinations. It also provided a more comprehensive understanding of model-generated outputs, ensuring the detection of hallucinations was both precise and insightful. The hyperparameters for fine-tuning are comprehensively detailed in Table 5.18.

The results are shown in Table 5.19

| Hyperparameter   | Value                          |
|------------------|--------------------------------|
| train dataloader | validation set (1,000 samples) |
| evaluator        | trial set (80 samples)         |
| epochs           | 5                              |
| evaluation steps | 10,000                         |
| warm-up steps    | 10% of train data for warm-up  |

Table 5.18: Hyperparameters used for the hallucination detection model fine-tuning

| <b>acc</b><br><b>test-agn</b> | <b>rho</b><br><b>test-agn</b> | <b>acc</b><br><b>test-aware</b> | <b>acc</b><br><b>test-aware</b> | <b>trial</b> |
|-------------------------------|-------------------------------|---------------------------------|---------------------------------|--------------|
| 0.778                         | 0.668                         | 0.795                           | 0.685                           | 0.85         |

Table 5.19: Fine-tune hal-detect model’s results

## 5.6 Voting Classifier

In our final approach, we adopted an ensemble technique known as the Voting Classifier to improve the robustness and accuracy of our predictive model. The Voting Classifier operates on the principle of aggregating the outputs of multiple individual classifiers to arrive at a final decision. Each classifier within the ensemble generates its prediction for the output class based on the given input data, and the Voting Classifier consolidates these predictions. The final predicted class is determined by a majority vote, whereby the class that receives the highest number of votes from the individual classifiers is selected as the final output.

The strength of this ensemble approach lies in its ability to harness the unique advantages of each constituent classifier. In our case, we incorporated several models, from previous chapters. Specifically, we utilized the pre-trained hallucination detection model, its fine-tuned version and the fine-tuned NLI model, each of which was trained to address the task from different perspectives and through varying methodologies. By combining these diverse models, we could capitalize on each approach’s individual strengths. Certain classifiers may be better suited to specific aspects or characteristics of the data; thus, their predictions reflect these nuances. However, by allowing multiple models to contribute to the decision-making process, the ensemble helps to balance out any potential weaknesses or biases that may arise when relying on a single model.

Moreover, the Voting Classifier serves as an effective strategy for mitigating the risks of overfitting or bias that may be present in individual models. While any single model may exhibit limitations due to noise or variability within the dataset, the ensemble method distributes the predictive responsibility across multiple classifiers, thus reducing the impact of any particular model’s deficiencies. This approach is particularly valuable when dealing with complex and heterogeneous datasets, as it increases the likelihood that diverse patterns within the data are captured and accurately modeled.

Through its collective decision-making framework, the ensemble technique enhances the overall predictive performance of the system. By pooling the predictions of various classifiers, the Voting Classifier provides a more reliable and consistent prediction than any single model could achieve on its own. This approach proved especially effective in addressing our dataset’s inherent complexity and variability, contributing to a more nuanced and accurate understanding of the underlying phenomena. The Voting Classifier allowed us to navigate the intricacies of the data better, ensuring that the final predictions were well-rounded and representative of the multiple factors at play, ultimately leading to improved model generalization and predictive accuracy.

The results of this method are shown in Table 5.20, which achieved the higher accuracy of all of our submissions for the Semeval competition.

| Method                                | acc.↑        | rho↑         |
|---------------------------------------|--------------|--------------|
| <b>Model-aware</b>                    |              |              |
| Baseline Model                        | 0.745        | 0.488        |
| Fine-tune hal-detect model            | 0.795        | 0.685        |
| NLI model                             | 0.77         | 0.591        |
| Voting Classifier-majority vote       | <b>0.799</b> | 0.691        |
| Voting Classifier-averaged percentage | <b>0.799</b> | <b>0.693</b> |
| <b>Model-agnostic</b>                 |              |              |
| Baseline Model                        | 0.697        | 0.402        |
| Fine-tune hal-detect model            | 0.778        | 0.668        |
| NLI model                             | 0.751        | 0.548        |
| Voting Classifier-majority vote       | <b>0.78</b>  | 0.632        |
| Voting Classifier-averaged percentage | <b>0.78</b>  | <b>0.643</b> |

Table 5.20: Final results for model-aware and model-agnostic variants. **Bold** denotes best results. The two Voting Classifiers differentiate from the method applied to calculate the p(‘Hallucination)

## 5.7 Prompting

As previously discussed, our final approach centered on applying Prompt Engineering. To initiate this method, we designed and implemented a zero-shot prompt, deploying it across a range of models, including LLaMA-2 (7B, 13B, and 70B), LLaMA-3 70B, and GEMMA, among others. Following this, we extended our approach to include a few-shot prompt and further explored the Chain-of-Thought (CoT) reasoning by incorporating the prompt illustrated in Table 5.24. The outcomes of each respective prompting method—zero-shot, few-shot, and CoT—are systematically presented in Table 5.21, Table 5.22, and Table 5.23, respectively. These results offer a comparative analysis of the performance across different models and prompting techniques.

| Model       | Acc test-agnostic | Acc test-aware |
|-------------|-------------------|----------------|
| Llama-2 7b  | 0.60              | 0.62           |
| Llama-2 13b | 0.64              | 0.67           |
| Llama-2 70b | <b>0.77</b>       | <b>0.75</b>    |
| Llama-3 70b | <b>0.77</b>       | <b>0.75</b>    |
| Gemma 2b    | 0.60              | 0.61           |
| Gemma 7b    | 0.74              | 0.7            |
| Mistral 7b  | 0.46              | 0.47           |

Table 5.21: Results for Zero-shot

| Model       | Acc test-agnostic | Acc test-aware |
|-------------|-------------------|----------------|
| Llama-2 7b  | **                | **             |
| Llama-2 13b | 0.42              | 0.51           |
| Llama-2 70b | 0.62              | 0.64           |
| Llama-3 8b  | **                | **             |
| Llama-3 70b | 0.62              | 0.65           |
| Gemma 2b    | 0.57              | 0.55           |
| Gemma 7b    | **                | **             |
| Mistral 7b  | 0.5               | 0.54           |

Table 5.22: Results for Few-shot

| Model       | Acc test-agnostic | Acc test-aware |
|-------------|-------------------|----------------|
| Llama-2 7b  | 0.63              | 0.65           |
| Llama-2 13b | 0.58              | 0.56           |
| Llama-2 70b | 0.57              | 0.61           |
| Llama-3 8b  | **                | **             |
| Llama-3 70b | 0.58              | 0.62           |
| Gemma 2b    | 0.59              | 0.63           |
| Gemma 7b    | 0.60              | 0.63           |
| Mistral 7b  | 0.49              | 0.46           |

Table 5.23: Results for CoT

**Observations:** The results from our experiments indicate that the most effective model in this context is LLaMA-2 70B, particularly when utilizing the zero-shot prompting approach. This superior performance

| Method              | Prompt                                                                                                                                                                                                                                                                                                                                                              |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Zero-shot Prompting | [Sentence 1]: Freedom is worthless.<br>[Sentence 2]: Freedom is invaluable.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.                                                                                                                                                                                                             |
| Few-shot            | In the following example, the Sentence 1 is supported by the Sentence 2:<br>example_1<br>In the following example, the Sentence 1 is not supported by the Sentence 2:<br>example_2<br>Now, it is your turn to evaluate the following example.<br>[Sentence 1]: hyp<br>[Sentence 2]: tgt<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No. |
| Chain-of-Thought    | [Sentence 1]: Freedom is worthless.<br>[Sentence 2]: Freedom is invaluable.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>Explain your answer.                                                                                                                                                                                     |

Table 5.24: Prompts used for each method

can be attributed to the model’s advanced architecture, coupled with the prompt’s simplification, enabling more precise input processing. Additionally, we observed a tendency in smaller models, such as LLaMA-7B, to exhibit biased behavior by consistently assigning a single label to all data points. Specifically, the model would classify every instance as either "Hallucination" or "Not Hallucination," depending on its training configuration, leading to inflated recall rates, as shown in Table 5.25.

| Model       | Recall test-agnostic<br>Hallucination/Not<br>Hallucination | Recall test-aware<br>Hallucination/Not<br>Hallucination |
|-------------|------------------------------------------------------------|---------------------------------------------------------|
| Llama-2 7b  | 0.25/0.85                                                  | 0.39/0.76                                               |
| Llama-2 13b | 0.29/0.82                                                  | 0.29/0.82                                               |
| Llama-2 70b | 0.5/0.95                                                   | 0.52/0.88                                               |
| Llama-3 70b | 0.53/0.93                                                  | 0.41/0.95                                               |
| Gemma 2b    | 0.04/0.99                                                  | 0.03/0.96                                               |
| Gemma 7b    | 0.49/0.91                                                  | 0.42/0.87                                               |
| Mistral 7b  | 0.5/0.93                                                   | 0.4/0.8                                                 |

Table 5.25: Recall for each dataset of zero-shot prompt

Moreover, our findings suggest that as the complexity of the prompt increased, the model’s comprehension and

overall accuracy significantly declined. This highlights the sensitivity of these models to prompt complexity, with overly intricate prompts leading to reduced performance and diminished understanding of the task at hand. Furthermore, in Table 5.22 and Table 5.23, the entries marked with \*\* indicate cases where the model’s responses were entirely irrelevant to the input, making it impossible to extract a valid answer. A couple of completely irrelevant responses and the models that produced them are shown in Table 5.26

| Model      | Input                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Output                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Llama-3 8b | "[Sentence 1]: Tom thought Mary probably wouldn't give up French.<br>[Sentence 2]: Tom thought Mary would probably fail her French test.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>Explain your answer.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <a href="https://www.duolingo.com/bookmark/1206a4">https://www.duolingo.com/bookmark/1206a4</a>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| gemma 7b   | First, you will be provided with 2 examples with the correct answer, similar to the question you will need to answer.<br>[Sentence 1]: Freedom is worthless.<br>[Sentence 2]: Freedom is invaluable.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>[Correct Answer]: No, Sentence 1 is not supported by the Sentence 2.<br>[Sentence 1]: This isn't your house anymore.<br>[Sentence 2]: You don't live here anymore.<br>Is the Sentence 1 supported by the Sentence 2?<br>Answer with Yes or No.<br>[Correct Answer]: Yes, Sentence 1 is supported by the Sentence 2.<br>Now, it is your turn to evaluate the following example<br>[Sentence 1]: Can anyone corroborate that?<br>[Sentence 2]: Can anyone back you up on that? Sentence 1 is not supported by the Sentence 2 Answer with Yes or No. | ## The answers for each sentence:<br>**Freedom and Can Anyone Corroborate/Back Me Up?** ###<br>Sentences about freedom : (sentence one) - FREEDOM IS WORTHLESS / (sentence two)-FREEDOM HAS VALUE . The sentences are saying different things despite using seemingly synonymous phrases like 'freedom', valuable & useless'. They differ in tone of voice , purpose as well context which significantly impacts their meanings even though they both deal broadly around concepts related directly opposite from valuedness versus worthlessness.. Therefore I believe these statements do NOT carry identical semantic weight hence my response was NO.**_***sentences regarding living space:** - [Sentece One] THIS AIN2019T YOUR HOUSE ANYMORE./-[Sentence Two]-YOU DON'T LIVE HERE AGAIN AND THE ANSWER FOR BOTH SENTENCES ARE YES BECAUSE THEY CARRY SAME MEANING BUT JUST WITH DIFFERENT WAYS OF SAYINGS THAT STILL HAVE A LIKE TO ONE Another. |

Table 5.26: Outputs completely irrelevant to input

In our final effort to enhance performance, we focused on refining the zero-shot prompts used in LLaMA-2 70B, which had previously demonstrated the highest accuracy across tasks. For each specific task, we systematically adjusted the prompts to better align with the task requirements. The modified prompts for each task are detailed in Table 5.27. The results after these adjustments are provided in Table 5.28, which once again demonstrates that the highest results come from simple prompts.

| Task | Prompt                                                                                                                                                                                                                                                                                     |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DM   | [Sentence 1]: src<br>[Sentence 2]: hyp<br>[Sentence 3]: tgt<br>A machine was given the task of giving<br>the definition of Sentence 1. The<br>machine’s answer was Sentence 2 while<br>the expected answer was Sentence 3.<br>Did the machine answer correctly?<br>Answer with Yes or No.  |
| PG   | [Sentence 1]: src<br>[Sentence 2]: hyp<br>A machine was given the task of giving<br>a paraphrase of Sentence 1. The<br>machine’s answer was Sentence 2<br>Did the machine answer correctly?<br>Answer with Yes or No.                                                                      |
| MT   | [Sentence 1]: src<br>[Sentence 2]: hyp<br>[Sentence 3]: tgt<br>A machine was given the task of giving<br>the translation of Sentence 1. The<br>machine’s answer was Sentence 2 while<br>the expected answer was Sentence 3.<br>Did the machine answer correctly?<br>Answer with Yes or No. |

Table 5.27: Prompts used for each Task

| Model       | Accuracy test-agnostic | Accuracy test-aware |
|-------------|------------------------|---------------------|
| Llama-3 70b | 0.66                   | 0.63                |

Table 5.28: Results from task separation prompt

## 5.8 Overall Results

Following the completion of all experiments, the results are summarized in Table 5.29, which presents the highest accuracy achieved by each method.

| Method                      | Model-Aware Test Set | Model-Agnostic Test Set |
|-----------------------------|----------------------|-------------------------|
| Baseline System             | 0.75                 | 0.70                    |
| Sentence Similarity         | 0.75                 | 0.62                    |
| COMET                       | 0.73                 | 0.74                    |
| NLI Model                   | 0.75                 | 0.77                    |
| Vectara Hallucination Model | 0.77                 | 0.79                    |
| Voting Classifier           | <b>0.78</b>          | <b>0.80</b>             |
| Prompting (Zero-Shot)       | 0.77                 | 0.75                    |

Table 5.29: Accuracy results for all methods on the Model-Aware and Model-Agnostic test sets.

The highest accuracy was achieved using the Voting Classifier method, which also represented our top submission for the SemEval-2024 Task 6 competition [16].

While the accuracy metric provides an indication of the proportion of correct predictions, it does not offer insights into the nature of the errors or their underlying causes. To address this, we present an analysis of the errors associated with each method, categorized by label.

In Figure 5.8.1, the bar plots illustrate the percentage of errors for each label, calculated relative to the initial number of samples per label in both the model-aware and model-agnostic test sets.

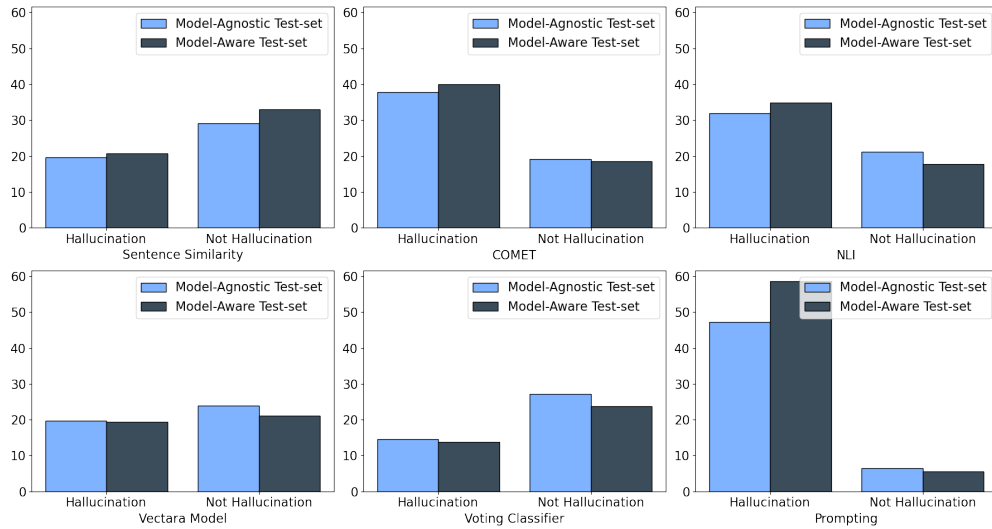


Figure 5.8.1: Percentage of errors per label in the model-agnostic and aware test set.

In Figure 5.8.2, the bar plots illustrate the percentage of errors for each label, calculated relative to the initial number of samples per task in both the model-aware and model-agnostic test sets.

In Figure 5.8.3, the bar plots illustrate the percentage of errors for each label, calculated relative to the initial number of samples per p('Hallucination') in both the model-aware and model-agnostic test sets.

We observe that the Paraphrase Generation task demonstrates superior performance compared to the other tasks, which is expected given the nature of paraphrasing. In contrast, the Definition Modeling task exhibits

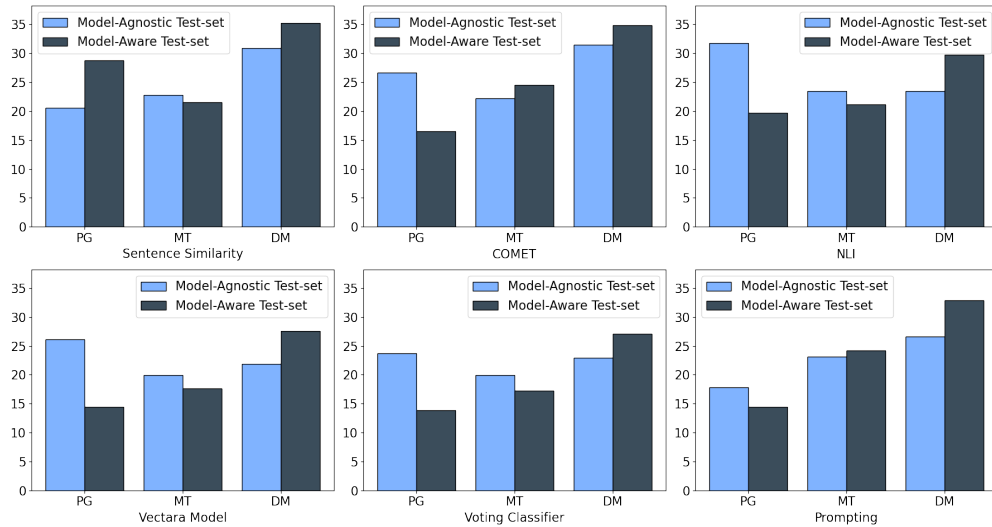
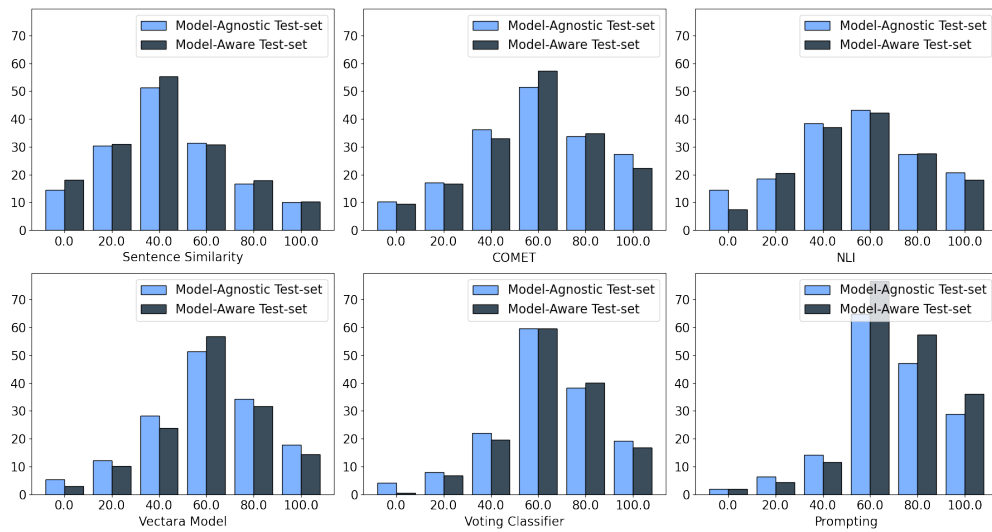


Figure 5.8.2: Percentage of errors per task in the model-agnostic and aware test set.

Figure 5.8.3: Percentage of errors per  $p(\text{'Hallucination'})$  in the model-agnostic and aware test set.

significantly lower accuracy, which is also reasonable due to the inherent difficulty associated with generating accurate definitions.

An important observation regarding  $p(\text{Hallucination})$  is the marked disparity in error rates when  $p(\text{Hallucination})$  equals 0.6 or 0.4. This variation can be attributed to the manner in which  $p(\text{Hallucination})$  is computed, reflecting the level of agreement among the five annotators. The discrepancies indicate that even among multiple annotators, consensus on whether a case constitutes a hallucination is not always achieved.



# Chapter 6

## Conclusion

This thesis has addressed the critical issue of hallucinations in Large Language Models (LLMs), particularly in the context of Natural Language Generation (NLG) tasks, where such errors can undermine the reliability of NLP systems. The widespread adoption of LLMs across various domains has heightened the importance of developing robust methods to detect and mitigate hallucinations, especially in black-box settings where the inner workings of the models remain opaque. Through participation in the SemEval-2024 Task 6 (SHROOM), this research has explored novel approaches to hallucination detection that apply to challenging domains such as Machine Translation, Definition Modelling, and Paraphrase Generation.

Key contributions of this work include the fine-tuning of pre-trained models on hallucination detection and Natural Language Inference datasets, achieving superior performance compared to baseline models. The integration of these models into a Voting Classifier framework has demonstrated competitive accuracy in detecting hallucinations across model-aware and model-agnostic test sets. Additionally, this research has introduced a novel prompting approach to enhance interpretability, guiding models to identify semantic inconsistencies in the generated outputs. By employing sentence similarity metrics, the methodology has further enabled a quantitative assessment of hallucinations, grounded in the comparison between generated and input texts.

The results achieved in this thesis underscore the potential of combining multiple strategies, including fine-tuning, ensemble methods, and semantic metrics, to improve hallucination detection accuracy. The analysis of error patterns across different tasks has also provided valuable insights into the nature of hallucinations, revealing the varying degrees of challenge posed by different types of NLG tasks. For instance, the superior performance observed in Paraphrase Generation is consistent with the task's inherent properties, whereas the lower accuracy in Definition Modeling reflects the greater difficulty of accurately modeling definitions. Furthermore, the study of annotation disagreement in  $p(\text{Hallucination})$  scores has highlighted the complexities involved in achieving consensus on the presence of hallucinations, suggesting areas for further refinement in evaluation protocols.

In conclusion, this research advances the state of the art in hallucination detection for LLMs by proposing a range of methods that are both effective and computationally efficient within a black-box setup. The findings lay a foundation for future work aimed at enhancing the reliability of LLM-based systems by addressing the intrinsic challenge of hallucinations. Future directions could involve refining annotation strategies, exploring hybrid detection methods that combine model-internal and external evidence, and expanding the evaluation to a wider array of NLG tasks and model architectures. By continuing to investigate the interplay between task-specific factors, model design, and detection techniques, the NLP community can move closer to building more trustworthy language generation systems.

### Future Work

1. **Task-Specific Techniques:** Develop specialized methods tailored to the characteristics of each Natural Language Generation (NLG) task. By considering the unique nature and requirements of tasks

such as summarization, translation, and dialogue generation, it is possible to create more effective hallucination detection approaches.

2. **Multi-Modal Approaches:** Expand the scope of research to include other domains, such as visual question answering and image captioning. Exploring hallucination detection in multi-modal tasks, where text is generated in conjunction with other data modalities, could lead to more comprehensive and robust methods.
3. **Integration of External Knowledge Bases:** Investigate the use of Retrieval-Augmented Generation (RAG) and other techniques to enhance the verification of factual accuracy in generated content. Leveraging external knowledge bases, such as Wikidata or domain-specific databases, could help validate the truthfulness of generated outputs.
4. **Explainable AI (XAI) Techniques:** Develop explainability methods to gain deeper insights into the nature of hallucinations and understand the underlying reasons for their occurrence. Techniques such as attention visualization, counterfactual analysis, and interpretable models could help elucidate the sources of errors in generated text.
5. **Automatic Hallucination Correction:** Beyond detecting hallucinations, explore methods to automatically correct or revise model outputs to make them factually consistent.

# Chapter 7

## Bibliography

- [1] *Accuracy vs. precision vs. recall in machine learning: what's the difference?* <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall>. URL:
- [2] Azaria, A. and Mitchell, T. *The Internal State of an LLM Knows When It's Lying*. 2023. arXiv: [2304.13734 \[cs.CL\]](#).
- [3] Ba, J. L., Kiros, J. R., and Hinton, G. E. *Layer Normalization*. 2016. arXiv: [1607.06450 \[stat.ML\]](#). URL:
- [4] Bengio, Y., Ducharme, R., and Vincent, P. “A Neural Probabilistic Language Model”. In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL:
- [5] Brown, T. B. et al. *Language Models are Few-Shot Learners*. 2020. arXiv: [2005.14165 \[cs.CL\]](#). URL:
- [6] Cao, Z., Yang, Y., and Zhao, H. “AutoHall: Automated Hallucination Dataset Generation for Large Language Models”. In: *ArXiv abs/2310.00259* (2023). URL:
- [7] Chen, X. et al. “FactCHD: Benchmarking Fact-Conflicting Hallucination Detection”. In: *ArXiv abs/2310.12086* (2023). URL:
- [8] Chern, I.-C. et al. *FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios*. 2023. arXiv: [2307.13528 \[cs.CL\]](#).
- [9] DataScientest. *Pearson and Spearman Correlations: A Guide to Understanding and Applying Correlation Methods*. Accessed: 2024-09-29. 2024. URL:
- [10] Devlin, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: [1810.04805 \[cs.CL\]](#). URL:
- [11] Dhuliawala, S. et al. *Chain-of-Verification Reduces Hallucination in Large Language Models*. 2023. arXiv: [2309.11495 \[cs.CL\]](#).
- [12] Duan, H., Yang, Y., and Tam, K. Y. “Do LLMs Know about Hallucination? An Empirical Investigation of LLM’s Hidden States”. In: 2024. URL:
- [13] Dziri, N. et al. “Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark”. In: *Transactions of the Association for Computational Linguistics* 10 (Sept. 2022), pp. 1066–1083. ISSN: 2307-387X. DOI: [10.1162/tac1\\_a\\_00506](#). eprint: URL:
- [14] Filandrianos, G. et al. “Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors”. In: *arXiv preprint arXiv:2305.17055* (2023).
- [15] Giadikiaroglou, P. et al. *Puzzle Solving using Reasoning of Large Language Models: A Survey*. 2024. arXiv: [2402.11291 \[cs.CL\]](#). URL:
- [16] Grigoriadou, N. et al. “AILS-NTUA at SemEval-2024 Task 6: Efficient model tuning for hallucination detection and analysis”. In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by A. K. Ojha et al. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1549–1560. DOI: [10.18653/v1/2024.semeval-1.222](#). URL:
- [17] Guerreiro, N. M., Voita, E., and Martins, A. “Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by A. Vlachos and I.

- Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1059–1075. DOI: [10.18653/v1/2023.eacl-main.75](https://doi.org/10.18653/v1/2023.eacl-main.75). URL:
- [18] He, K. et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [19] Huang, L. et al. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: [2311.05232](https://arxiv.org/abs/2311.05232) [cs.CL].
- [20] Ji, Z. et al. “Survey of Hallucination in Natural Language Generation”. In: *ACM Comput. Surv.* 55.12 (Mar. 2023). ISSN: 0360-0300. DOI: [10.1145/3571730](https://doi.org/10.1145/3571730). URL:
- [21] Ji, Z. et al. “Survey of Hallucination in Natural Language Generation”. In: *ACM Comput. Surv.* 55.12 (Mar. 2023). ISSN: 0360-0300. DOI: [10.1145/3571730](https://doi.org/10.1145/3571730). URL:
- [22] Jiang, A. Q. et al. *Mistral 7B*. 2023. arXiv: [2310.06825](https://arxiv.org/abs/2310.06825) [cs.CL]. URL:
- [23] Kadavath, S. et al. *Language Models (Mostly) Know What They Know*. 2022. arXiv: [2207.05221](https://arxiv.org/abs/2207.05221) [cs.CL].
- [24] Lee, N. et al. *Factuality Enhanced Language Models for Open-Ended Text Generation*. 2023. arXiv: [2206.04624](https://arxiv.org/abs/2206.04624) [cs.CL].
- [25] Lewis, M. et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: [1910.13461](https://arxiv.org/abs/1910.13461) [cs.CL]. URL:
- [26] Li, J. et al. “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. DOI: [10.18653/v1/2023.emnlp-main.397](https://doi.org/10.18653/v1/2023.emnlp-main.397). URL:
- [27] Li, J. et al. “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. DOI: [10.18653/v1/2023.emnlp-main.397](https://doi.org/10.18653/v1/2023.emnlp-main.397). URL:
- [28] Li, J. et al. *The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models*. 2024. arXiv: [2401.03205](https://arxiv.org/abs/2401.03205) [cs.CL].
- [29] Lin, S., Hilton, J., and Evans, O. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. 2022. arXiv: [2109.07958](https://arxiv.org/abs/2109.07958) [cs.CL].
- [30] Liu, T. et al. “A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6723–6737. DOI: [10.18653/v1/2022.acl-long.464](https://doi.org/10.18653/v1/2022.acl-long.464). URL:
- [31] Lymperopoulos, D. et al. *Optimal and efficient text counterfactuals using Graph Neural Networks*. 2024. arXiv: [2408.01969](https://arxiv.org/abs/2408.01969) [cs.CL]. URL:
- [32] Manakul, P., Liusie, A., and Gales, M. “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9004–9017. DOI: [10.18653/v1/2023.emnlp-main.557](https://doi.org/10.18653/v1/2023.emnlp-main.557). URL:
- [33] Manakul, P., Liusie, A., and Gales, M. J. F. *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*. 2023. arXiv: [2303.08896](https://arxiv.org/abs/2303.08896) [cs.CL].
- [34] Maynez, J. et al. “On Faithfulness and Factuality in Abstractive Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 1906–1919. DOI: [10.18653/v1/2020.acl-main.173](https://doi.org/10.18653/v1/2020.acl-main.173). URL:
- [35] Mickus, T. et al. “SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes”. In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, June 2024.
- [36] Mikolov, T. et al. “Recurrent neural network based language model”. In: vol. 2. Sept. 2010, pp. 1045–1048. DOI: [10.21437/Interspeech.2010-343](https://doi.org/10.21437/Interspeech.2010-343).
- [37] Min, S. et al. *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. 2023. arXiv: [2305.14251](https://arxiv.org/abs/2305.14251) [cs.CL].

- 
- [38] Muhlgay, D. et al. *Generating Benchmarks for Factuality Evaluation of Language Models*. 2024. arXiv: [2307.06908](#) [cs.CL].
- [39] Panagiotopoulos, I. et al. *AILS-NTUA at SemEval-2024 Task 9: Cracking Brain Teasers: Transformer Models for Lateral Thinking Puzzles*. 2024. arXiv: [2404.01084](#) [cs.CL]. URL:
- [40] Panagiotopoulos, I. et al. *RISCORE: Enhancing In-Context Riddle Solving in Language Models through Context-Reconstructed Example Augmentation*. 2024. arXiv: [2409.16383](#) [cs.CL]. URL:
- [41] Radford, A. et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [42] Raffel, C. et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. arXiv: [1910.10683](#) [cs.LG]. URL:
- [43] Raunak, V., Menezes, A., and Junczys-Dowmunt, M. “The Curious Case of Hallucinations in Neural Machine Translation”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 1172–1183. DOI: [10.18653/v1/2021.naacl-main.92](#). URL:
- [44] Rawte, V., Sheth, A., and Das, A. *A Survey of Hallucination in Large Foundation Models*. 2023. arXiv: [2309.05922](#) [cs.AI].
- [45] Rei, R. et al. “COMET: A Neural Framework for MT Evaluation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 2685–2702. DOI: [10.18653/v1/2020.emnlp-main.213](#). URL:
- [46] Rohrbach, A. et al. “Object Hallucination in Image Captioning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by E. Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4035–4045. DOI: [10.18653/v1/D18-1437](#). URL:
- [47] Thoppilan, R. et al. *LaMDA: Language Models for Dialog Applications*. 2022. arXiv: [2201.08239](#) [cs.CL]. URL:
- [48] Vaswani, A. et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: [1706.03762](#). URL:
- [49] Vinyals, O. and Le, Q. “A Neural Conversational Model”. In: *arXiv preprint arXiv:1506.05869* (2015). URL:
- [50] Workshop, B. et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 2023. arXiv: [2211.05100](#) [cs.CL]. URL:
- [51] Xu, Z., Jain, S., and Kankanhalli, M. S. “Hallucination is Inevitable: An Innate Limitation of Large Language Models”. In: *ArXiv* abs/2401.11817 (2024). URL:
- [52] Ye, H. et al. *Cognitive Mirage: A Review of Hallucinations in Large Language Models*. 2023. arXiv: [2309.06794](#) [cs.CL].
- [53] Zhang, Y. et al. “Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models”. In: *ArXiv* abs/2309.01219 (2023). URL:
- [54] Zhou, C. et al. “Detecting Hallucinated Content in Conditional Neural Sequence Generation”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by C. Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 1393–1404. DOI: [10.18653/v1/2021.findings-acl.120](#). URL:
-



## Appendix A

### Thresholds Used for Sentence Similarity

| Model                                                                | Threshold<br>Euclidean | Threshold<br>Cosine<br>Similarity | Threshold<br>Manhattan |
|----------------------------------------------------------------------|------------------------|-----------------------------------|------------------------|
| bert-base-uncased                                                    | 0.83                   | 6.96                              | 151.74                 |
| stsb-roberta-large                                                   | 0.7                    | 20.47                             | 514.38                 |
| all-distilroberta-v1                                                 | 0.61                   | 0.88                              | 19.21                  |
| all-MiniLM-L6-v2                                                     | 0.63                   | 0.86                              | 13.39                  |
| all-MiniLM-L12-v2                                                    | 0.63                   | 0.86                              | 13.37                  |
| all-mpnet-base-v2                                                    | 0.65                   | 0.84                              | 18.33                  |
| all-roberta-large-v1                                                 | 0.65                   | 0.83                              | 21.14                  |
| paraphrase-MiniLM-L6-v2                                              | 0.66                   | 5.36                              | 83.37                  |
| paraphrase-mpnet-base-v2                                             | 0.7                    | 2.52                              | 54.28                  |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.68                   | 0.8                               | 12.4                   |
| paraphrase-MiniLM-L12                                                | 0.68                   | 4.1                               | 63.68                  |
| paraphrase-MiniLM-L3                                                 | 0.62                   | 3.74                              | 57.87                  |
| paraphrase-TinyBERT-L6                                               | 0.65                   | 5.2                               | 114.91                 |
| paraphrase-distilroberta-base                                        | 0.66                   | 13.16                             | 288.61                 |
| xlm-r-distilroberta                                                  | 0.69                   | 5.17                              | 111.42                 |
| stsb-roberta-base-v2                                                 | 0.67                   | 14.08                             | 307.57                 |
| stsb-roberta-large                                                   | 0.7                    | 20.47                             | 514.38                 |
| stsb-distilroberta-base-v2                                           | 0.67                   | 13.84                             | 301.1                  |
| stsb-mpnet-base-v2                                                   | 0.69                   | 2.63                              | 56.52                  |
| multi-qa-distilbert-cos-v1                                           | 0.65                   | 0.83                              | 18.34                  |
| multi-qa-MiniLM-L6-cos-v1                                            | 0.68                   | 0.8                               | 12.4                   |
| sentence-t5-base                                                     | 0.9                    | 0.45                              | 9.85                   |
| sentence-t5-large                                                    | 0.89                   | 0.48                              | 10.6                   |
| gtr-t5-base                                                          | 0.78                   | 0.66                              | 14.67                  |
| gtr-t5-large                                                         | 0.8                    | 0.64                              | 14.08                  |
| nq-distilbert-base-v1                                                | 0.59                   | 12.28                             | 271.32                 |
| nli-distilroberta-base-v2                                            | 0.7                    | 12.5                              | 270.87                 |
| sentence-transformers_distilbert-multilingual-nli-stsb-quora-ranking | 0.91                   | 4.19                              | 92.43                  |
| sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2          | 0.72                   | 3.63                              | 56.32                  |
| sentence-transformers_paraphrase-multilingual-mpnet-base-v2          | 0.76                   | 2.17                              | 46.39                  |
| sentence-transformers_distiluse-base-multilingual-cased-v2           | 0.7                    | 0.75                              | 13.35                  |
| sentence-transformers_paraphrase-xlm-r-multilingual-v1               | 0.69                   | 5.17                              | 111.42                 |
| sentence-transformers-quora-distilbert-multilingual                  | 0.91                   | 4.19                              | 92.43                  |
| sentence-transformers_stsb-xlm-r-multilingual                        | 0.75                   | 10.74                             | 234.54                 |
| sentence-transformers_distiluse-base-multilingual-cased              | 0.7                    | 0.75                              | 13.35                  |

Table A.1: Thresholds used for model-agnostic test set

| Model                                                                 | Threshold<br>Euclidean | Threshold<br>Cosine<br>Similarity | Threshold<br>Manhattan |
|-----------------------------------------------------------------------|------------------------|-----------------------------------|------------------------|
| bert-base-uncased                                                     | 0.82                   | 7.78                              | 170.22                 |
| stsb-roberta-large                                                    | 0.61                   | 23.86                             | 601.39                 |
| all-distilroberta-v1                                                  | 0.57                   | 0.93                              | 20.27                  |
| all-MiniLM-L6-v2                                                      | 0.6                    | 0.9                               | 13.99                  |
| all-MiniLM-L12-v2                                                     | 0.6                    | 0.9                               | 13.93                  |
| all-mpnet-base-v2                                                     | 0.62                   | 0.87                              | 18.75                  |
| all-roberta-large-v1                                                  | 0.58                   | 0.92                              | 23.43                  |
| paraphrase-MiniLM-L6-v2                                               | 0.6                    | 5.6                               | 86.48                  |
| paraphrase-mpnet-base-v2                                              | 0.64                   | 2.8                               | 60.36                  |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.6                    | 0.89                              | 13.83                  |
| paraphrase-MiniLM-L12                                                 | 0.59                   | 4.53                              | 70.71                  |
| paraphrase-MiniLM-L3                                                  | 0.57                   | 4.1                               | 64.03                  |
| paraphrase-TinyBERT-L6                                                | 0.6                    | 5.43                              | 120.02                 |
| paraphrase-distilroberta-base                                         | 0.58                   | 14.06                             | 306.93                 |
| xlm-r-distilroberta                                                   | 0.61                   | 5.47                              | 117.34                 |
| stsb-roberta-base-v2                                                  | 0.58                   | 16.71                             | 363.02                 |
| stsb-roberta-large                                                    | 0.61                   | 23.86                             | 601.39                 |
| stsb-distilroberta-base-v2                                            | 0.58                   | 15.16                             | 328.85                 |
| stsb-mpnet-base-v2                                                    | 0.59                   | 2.98                              | 64.07                  |
| multi-qa-distilbert-cos-v1                                            | 0.6                    | 0.9                               | 19.76                  |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.6                    | 0.89                              | 13.83                  |
| sentence-t5-base                                                      | 0.89                   | 0.48                              | 10.55                  |
| sentence-t5-large                                                     | 0.87                   | 0.52                              | 11.49                  |
| gtr-t5-base                                                           | 0.74                   | 0.73                              | 15.98                  |
| gtr-t5-large                                                          | 0.76                   | 0.69                              | 15.22                  |
| nq-distilbert-base-v1                                                 | 0.55                   | 13.11                             | 289.75                 |
| nli-distilroberta-base-v2                                             | 0.65                   | 13.5                              | 295.03                 |
| sentence-transformers__paraphrase-multilingual-MiniLM-L12-v2          | 0.64                   | 4.08                              | 63.35                  |
| sentence-transformers__paraphrase-multilingual-mpnet-base-v2          | 0.69                   | 2.42                              | 51.78                  |
| sentence-transformers__distiluse-base-multilingual-cased-v2           | 0.56                   | 0.91                              | 16.14                  |
| sentence-transformers__paraphrase-xlm-r-multilingual-v1               | 0.61                   | 5.47                              | 117.34                 |
| sentence-transformers__distilbert-multilingual-nli-stsb-quora-ranking | 0.88                   | 4.83                              | 106.69                 |
| sentence-transformers-quora-distilbert-multilingual                   | 0.88                   | 4.83                              | 106.69                 |
| sentence-transformers__stsb-xlm-r-multilingual                        | 0.67                   | 12.04                             | 265.35                 |
| sentence-transformers__distiluse-base-multilingual-cased              | 0.56                   | 0.91                              | 16.14                  |

Table A.2: Thresholds used for model-aware test set

| Model                                                                 | Threshold<br>Euclidean | Threshold<br>Cosine<br>Similarity | Threshold<br>Manhattan |
|-----------------------------------------------------------------------|------------------------|-----------------------------------|------------------------|
| bert-base-uncased                                                     | 0.82                   | 7.18                              | 156.9                  |
| stsb-roberta-large                                                    | 0.69                   | 21.01                             | 531.35                 |
| all-distilroberta-v1                                                  | 0.61                   | 0.89                              | 19.36                  |
| all-MiniLM-L6-v2                                                      | 0.63                   | 0.86                              | 13.36                  |
| all-MiniLM-L12-v2                                                     | 0.63                   | 0.86                              | 13.29                  |
| all-mpnet-base-v2                                                     | 0.65                   | 0.84                              | 18.18                  |
| all-roberta-large-v1                                                  | 0.66                   | 0.82                              | 20.83                  |
| paraphrase-MiniLM-L6-v2                                               | 0.64                   | 5.33                              | 82.99                  |
| paraphrase-mpnet-base-v2                                              | 0.68                   | 2.58                              | 55.66                  |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.68                   | 0.8                               | 12.43                  |
| paraphrase-MiniLM-L12                                                 | 0.66                   | 4.18                              | 64.48                  |
| paraphrase-MiniLM-L3                                                  | 0.61                   | 3.72                              | 57.97                  |
| paraphrase-TinyBERT-L6                                                | 0.65                   | 5.1                               | 113.03                 |
| paraphrase-distilroberta-base                                         | 0.66                   | 13.19                             | 286.01                 |
| xlm-r-distilroberta                                                   | 0.69                   | 5.07                              | 110.05                 |
| stsb-roberta-base-v2                                                  | 0.65                   | 14.65                             | 321.02                 |
| stsb-roberta-large                                                    | 0.69                   | 21.01                             | 531.35                 |
| stsb-distilroberta-base-v2                                            | 0.66                   | 13.98                             | 304.57                 |
| stsb-mpnet-base-v2                                                    | 0.67                   | 2.73                              | 59.05                  |
| multi-qa-distilbert-cos-v1                                            | 0.66                   | 0.82                              | 18.15                  |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.68                   | 0.8                               | 12.43                  |
| sentence-t5-base                                                      | 0.9                    | 0.45                              | 9.91                   |
| sentence-t5-large                                                     | 0.88                   | 0.49                              | 10.81                  |
| gtr-t5-base                                                           | 0.78                   | 0.66                              | 14.55                  |
| gtr-t5-large                                                          | 0.8                    | 0.64                              | 14.17                  |
| nq-distilbert-base-v1                                                 | 0.58                   | 12.43                             | 275.89                 |
| nli-distilroberta-base-v2                                             | 0.69                   | 12.7                              | 275.14                 |
| sentence-transformers__paraphrase-multilingual-MiniLM-L12-v2          | 0.7                    | 3.72                              | 57.37                  |
| sentence-transformers__paraphrase-multilingual-mpnet-base-v2          | 0.73                   | 2.25                              | 48.38                  |
| sentence-transformers__distiluse-base-multilingual-cased-v2           | 0.68                   | 0.78                              | 13.67                  |
| sentence-transformers__paraphrase-xlm-r-multilingual-v1               | 0.69                   | 5.07                              | 110.05                 |
| sentence-transformers__distilbert-multilingual-nli-stsb-quora-ranking | 0.91                   | 4.21                              | 93.05                  |
| sentence-transformers-quora-distilbert-multilingual                   | 0.91                   | 4.21                              | 93.05                  |
| sentence-transformers__stsb-xlm-r-multilingual                        | 0.74                   | 10.99                             | 242.3                  |
| sentence-transformers__distiluse-base-multilingual-cased              | 0.68                   | 0.78                              | 13.67                  |

Table A.3: Thresholds used for model-agnostic validation set

| Model                                                                 | Threshold<br>Euclidean | Threshold<br>Cosine<br>Similarity | Threshold<br>Manhattan |
|-----------------------------------------------------------------------|------------------------|-----------------------------------|------------------------|
| bert-base-uncased                                                     | 0.82                   | 7.7                               | 168.92                 |
| stsb-roberta-large                                                    | 0.62                   | 23.78                             | 601.14                 |
| all-distilroberta-v1                                                  | 0.58                   | 0.92                              | 20.07                  |
| all-MiniLM-L6-v2                                                      | 0.6                    | 0.89                              | 13.88                  |
| all-MiniLM-L12-v2                                                     | 0.61                   | 0.88                              | 13.58                  |
| all-mpnet-base-v2                                                     | 0.64                   | 0.85                              | 18.71                  |
| all-roberta-large-v1                                                  | 0.59                   | 0.91                              | 23.07                  |
| paraphrase-MiniLM-L6-v2                                               | 0.62                   | 5.42                              | 84.84                  |
| paraphrase-mpnet-base-v2                                              | 0.67                   | 2.68                              | 58.15                  |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.6                    | 0.89                              | 13.74                  |
| paraphrase-MiniLM-L12                                                 | 0.62                   | 4.38                              | 68.36                  |
| paraphrase-MiniLM-L3                                                  | 0.57                   | 4.04                              | 62.72                  |
| sentence-t5-base                                                      | 0.88                   | 0.48                              | 10.66                  |
| sentence-t5-large                                                     | 0.87                   | 0.51                              | 11.27                  |
| gtr-t5-base                                                           | 0.74                   | 0.73                              | 16.04                  |
| gtr-t5-large                                                          | 0.77                   | 0.68                              | 15.03                  |
| nq-distilbert-base-v1                                                 | 0.52                   | 13.04                             | 288.21                 |
| nli-distilroberta-base-v2                                             | 0.66                   | 13.01                             | 281.91                 |
| paraphrase-TinyBERT-L6                                                | 0.6                    | 5.34                              | 118.35                 |
| paraphrase-distilroberta-base                                         | 0.6                    | 13.79                             | 299.17                 |
| xlm-r-distilroberta                                                   | 0.62                   | 5.43                              | 117.06                 |
| stsb-roberta-base-v2                                                  | 0.6                    | 16.08                             | 351.81                 |
| stsb-roberta-large                                                    | 0.62                   | 23.78                             | 601.14                 |
| stsb-distilroberta-base-v2                                            | 0.6                    | 14.56                             | 313.91                 |
| stsb-mpnet-base-v2                                                    | 0.6                    | 2.89                              | 62.41                  |
| multi-qa-distilbert-cos-v1                                            | 0.59                   | 0.91                              | 20.06                  |
| multi-qa-MiniLM-L6-cos-v1                                             | 0.6                    | 0.89                              | 13.74                  |
| sentence-transformers__paraphrase-multilingual-MiniLM-L12-v2          | 0.66                   | 4.0                               | 61.87                  |
| sentence-transformers__paraphrase-multilingual-mpnet-base-v2          | 0.7                    | 2.35                              | 50.43                  |
| sentence-transformers__distiluse-base-multilingual-cased-v2           | 0.57                   | 0.91                              | 16.06                  |
| sentence-transformers__paraphrase-xlm-r-multilingual-v1               | 0.62                   | 5.43                              | 117.06                 |
| sentence-transformers__distilbert-multilingual-nli-stsb-quora-ranking | 0.89                   | 4.69                              | 102.91                 |
| sentence-transformers-quora-distilbert-multilingual                   | 0.89                   | 4.69                              | 102.91                 |
| sentence-transformers__stsb-xlm-r-multilingual                        | 0.69                   | 11.9                              | 260.67                 |
| sentence-transformers__distiluse-base-multilingual-cased              | 0.57                   | 0.91                              | 16.06                  |

Table A.4: Thresholds used for model-aware validation set

