Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

# Σχεδιασμός εικονικού βοηθού με χρήση Large Language Model (LLM) και εφαρμογή σε πλατφόρμα επενδύσεων ενεργειακής αποδοτικότητας

*Μελέτη και υλοποίηση*

---

Διπλωματικη Εργασια

του

**ΤΣΑΝΤΗΛΑ ΙΩΑΝΝΗ**

**Επιβλέπων:** Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

# Σχεδιασμός εικονικού βοηθού με χρήση Large Language Model (LLM) και εφαρμογή σε πλατφόρμα επενδύσεων ενεργειακής αποδοτικότητας

*Μελέτη και υλοποίηση*

---

Διπλωματικη Εργασια

του

**ΤΣΑΝΤΗΛΑ ΙΩΑΝΝΗ**

**Επιβλέπων:** Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3η Ιουλίου 2025.

| *(Υπογραφή)* | *(Υπογραφή)* | *(Υπογραφή)* |
|---|---|---|
| ..... | ..... | ..... |
| Δημήτριος Ασκούνης | Ιωάννης Ψαρράς | Βαγγέλης Μαρινάκης |
| Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. | Επικ. Καθηγητής Ε.Μ.Π. |

Αθήνα, Ιούλιος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων



*(Υπογραφή)*

.....
Τσαντήλας Ιωάννης

3 Ιουλίου 2025

*Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ*

*στους γονείς μου, Αναστασία και Βασίλειο*

# Περίληψη

Η ενεργειακή μετάβαση έχει γίνει προτεραιότητα για την Ευρωπαϊκή Ένωση. Για να επιτευχθεί, τίθενται στόχοι για την εξοικονόμηση ενέργειας, ενώ ενθαρρύνεται η ένταξη νέων τεχνολογιών στον ενεργειακό τομέα και η καινοτομία στις ενεργειακές υπηρεσίες. Στο πλαίσιο αυτό έχουν υλοποιηθεί πολλές πλατφόρμες και διαδικτυακές εφαρμογές, που έχουν ως στόχο την υποστήριξη των εμπλεκόμενων μερών στην ενεργειακή μετάβαση παρέχοντας ποικίλες υπηρεσίες σχετικές με την εξοικονόμηση ενέργειας, την ενεργειακή αποδοτικότητα, τις ανανεώσιμες πηγές ενέργειας, τη βιωσιμότητα κ.α. Ωστόσο, τέτοιες πλατφόρμες μπορεί συχνά να είναι δύσκολες στη χρήση, ειδικά όταν απευθύνονται σε κοινό μη τεχνολογικά καταρτισμένο, με αποτέλεσμα να περιορίζεται η εκμετάλλευσή τους. Η συγκέντρωση χρηστών από πολλές χώρες, διαφορετικών ηλικιών, με διαφορετικά επίπεδα εξοικείωσης με την τεχνολογία, συχνά δημιουργεί προκλήσεις. Συνεπώς, είναι σημαντικό να συμπεριληφθούν τα απαραίτητα εργαλεία για να βελτιώσουν την εμπειρία των χρηστών και να τους παρέχουν όλες τις απαραίτητες πληροφορίες για τη σωστή λειτουργία της εκάστοτε πλατφόρμας.

Στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία στοχεύει στον σχεδιασμό και την ανάπτυξη ενός Large Language Model (LLM) assistant, με δυνατότητα πιλοτικής εφαρμογής σε μία ερευνητική πλατφόρμα υποστήριξης επενδύσεων σε έργα ενεργειακής αποδοτικότητας σε κτίρια, ώστε να συλλεχθούν σχόλια από πραγματικούς χρήστες για την αξιολόγηση της αποτελεσματικότητάς του. Η πλατφόρμα διαθέτει ποικιλία υπηρεσιών και διαφορετικά είδη χρηστών. Διάφορα γλωσσικά μοντέλα (LLaMA, GPT-J, BLOOM) και knowledge structures θα εξεταστούν για τη δημιουργία και εκπαίδευση ενός πλήρως λειτουργικού βοηθού, ικανού να αλληλοεπιδρά με τους χρήστες και να τους καθοδηγεί στις λειτουργίες της πλατφόρμας. Ο βοηθός θα παρέχει επίσης αναλυτικές κατευθυντήριες γραμμές για την ολοκλήρωση των διάφορων λειτουργιών.

## Λέξεις Κλειδιά

Εικονικός Βοηθός (ΕΒ), Πλατφόρμα Επενδύσεων Ενεργειακής Αποδοτικότητας, Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ), Βελτιστοποίηση και Ρύθμιση ΜΓΜ

# Abstract

The energy transition has become a priority for the European Union. To achieve this, energy saving targets are set, while the integration of new technologies in the energy sector and innovation in energy services is encouraged. In this context, many platforms and web applications have been implemented to support stakeholders in the energy transition by providing a variety of services related to energy savings, energy efficiency, renewable energy, sustainability, etc. However, such platforms can often be difficult to use, especially when targeted at a non-technological audience, thus limiting their exploitation. The concentration of users from many countries, of different ages, with different levels of familiarity with technology, often creates challenges. It is therefore important to include the necessary tools to improve the user experience and provide users with all the information necessary for the proper functioning of the platform in question.

In this context, this thesis aims to design and develop a Large Language Model (LLM) assistant, with the possibility of piloting it on a research platform supporting investments in energy efficiency projects in buildings, in order to collect feedback from real users to evaluate its effectiveness. The platform has a variety of services and different types of users. Different language models (LLaMA, GPT-J, BLOOM) and knowledge structures will be explored to create and train a fully functional assistant capable of interacting with users and guiding them through the platform functions. The assistant will also provide detailed guidelines for completing the various functions.

## Keywords

Virtual Assistant (VA), Energy Efficiency Investment Platform, Large Language Models (LLMs), LLM Fine-tuning

# Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις ευχαριστίες μου σε όλους όσοι συνέβαλαν στην εκπόνησή της.

Καταρχάς, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Δημήτριο Ασκούνη για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο Εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης.

Ευχαριστώ θερμά επίσης τη διδακτορική φοιτήτρια Ντανιέλα Στογιάν για την καθοδήγησή της και την εξαιρετική συνεργασία που είχαμε. Η βοήθειά της στην εκπόνηση της εργασίας υπήρξε ανεκτίμητη, ενώ ο χαρακτήρας και η επαγγελματικότητά της υπήρξαν άψογα.

Ευχαριστώ ιδιαίτερα τους φίλους και τις φίλες μου, που στάθηκαν δίπλα μου όλα αυτά τα χρόνια, με τους οποίους αντιμετωπίσαμε μαζί τις δυσκολίες της φοίτησης στη ΣΗΜΜΥ. Ξεχωριστές ευχαριστίες απευθύνω σε έναν φίλο μου από το μακρινό Σαν Φρανσίσκο για την ανεκτίμητη βοήθειά του σε κρίσιμα σημεία της πορείας μου.

Ένα νεύμα εκτίμησης επίσης στον καφέ και την ευγενή του ουσία, την καφεΐνη, χωρίς την οποία οι αμέτρητες άυπνες νύχτες θα με είχαν καταβάλει.

Τέλος, και κυριότερα, θα ήθελα να εκφράσω τη βαθιά ευγνωμοσύνη μου στους γονείς μου, Βασίλειο και Αναστασία, καθώς και στα αδέλφια μου, Καλομοίρα, Βασιλική και Παναγιώτη, για τη στήριξη και την αγάπη που μου προσέφεραν όλα αυτά τα χρόνια.

*I did it!*

Αθήνα, Ιούλιος 2025

*Τσαντήλας Ιωάννης*

# Περιεχόμενα

**Περίληψη** **4**

**Abstract** **7**

**Ευχαριστίες** **9**

**1 Εισαγωγή** **19**
1.1 Η πρόκληση της ενεργειακής αναβάθμισης . . . . . 19
1.2 Ο ρόλος των ψηφιακών τεχνολογιών . . . . . 20
1.3 Η πλατφόρμα *ENERGATE* και ο στόχος της διπλωματικής εργασίας . . . . . 22
1.4 Οργάνωση του τόμου . . . . . 23

**I Θεωρητικό Μέρος** **25**

**2 Θεωρητικό υπόβαθρο** **27**
2.1 Τεχνητή Νοημοσύνη και Μηχανική Μάθηση . . . . . 27
2.1.1 Παραδείγματα μάθησης . . . . . 28
2.2 Λαργε Λανγυαγε Μοδελς . . . . . 28
2.2.1 Στόχοι προ-εκπαίδευσης . . . . . 29
2.2.2 Τεχνικές ευθυγράμμισης . . . . . 30
2.2.3 Κύκλος ζωής ΛΛΜ . . . . . 30
2.3 Μετρικές και σύνολα αξιολόγησης . . . . . 31
2.3.1 Οι μετρικές BLEU και ROUGE . . . . . 31
2.3.2 Σύνολα αξιολόγησης για την κατανόηση και τη συλλογιστική . . . . . 32
2.4 Ηθικοί και κοινωνικοί περιορισμοί . . . . . 33
2.4.1 Μεροληψία και δικαιοσύνη . . . . . 33
2.4.2 Ιδιωτικότητα και προστασία δεδομένων . . . . . 33
2.4.3 Περιβαλλοντική βιωσιμότητα . . . . . 34
2.4.4 Ασφάλεια, αξιοπιστία και συμμόρφωση με τους κανονισμούς . . . . . 34
2.5 Καινοτομία και διαφοροποίηση από προηγούμενες εργασίες . . . . . 34

**II Πρακτικό Μέρος** **37**

**3 Σχεδιασμός Μεθοδολογίας** **39**
3.1 Σύγκριση LLM με βάση τα Χαρακτηριστικά . . . . . 39

3.1.1 Επισκόπηση των Παραγόντων Σύγκρισης . . . . . 40
3.1.2 Επισκόπηση Μεγάλων Γλωσσικών Μοντέλων . . . . . 41
3.1.3 Συγκριτικός Πίνακας Χαρακτηριστικών . . . . . 50
3.1.4 Σχολιασμός Βαθμολογίας . . . . . 52
3.2 Σύγκριση LLM με βάση την Απόδοση . . . . . 53
3.2.1 Επισκόπηση του Open LLM Leaderboard . . . . . 53
3.2.2 Συγκριτικός Πίνακας Αποδόσεων . . . . . 55
3.2.3 Σχολιασμός Βαθμολογίας . . . . . 56

**4 Εφαρμογή Μεθοδολογίας** **57**
4.1 Πειραματική εφαρμογή LLaMA . . . . . 58
4.1.1 Τοπική ανάπτυξη μέσω Ollama . . . . . 58
4.1.2 Περιορισμοί υλικού και έρευνα κβαντοποίησης . . . . . 58
4.1.3 Υποδομή και οικονομικοί περιορισμοί . . . . . 59
4.1.4 Συμπεράσματα της τοπικής ανάπτυξης LLaMA . . . . . 59
4.2 Εφαρμογή στο Google Cloud Vertex AI . . . . . 60
4.2.1 Επιλογή πλατφόρμας και αρχική ρύθμιση . . . . . 60
4.2.2 Σύγκριση *Agent Builder* και *Fine-Tuning* . . . . . 60
4.2.3 Εποπτευόμενο fine-tuning του Gemini . . . . . 62
4.2.4 Μετατροπή της βάσης γνώσεων σε JSONL . . . . . 64
4.3 Εκτέλεση fine-tuning μοντέλου . . . . . 66
4.3.1 Εκτέλεση Fine-Tuning Εργασιών . . . . . 66
4.3.2 Παράμετροι εκπαίδευσης και έξοδοι μετρήσεων . . . . . 67
4.3.3 Άπληστη αναζήτηση υπερπαραμέτρων . . . . . 70

**5 Σχολιασμός Αποτελεσμάτων** **73**
5.1 Αρχική αξιολόγηση μοντέλου . . . . . 73
5.1.1 Δημιουργία του συνόλου δοκιμών . . . . . 73
5.1.2 Αλληλεπίδραση μοντέλου και συλλογή απαντήσεων . . . . . 74
5.1.3 Υπολογισμός Μετρήσεων . . . . . 75
5.2 Ενίσχυση μοντέλου με αναγνώριση τοποθεσίας . . . . . 77
5.2.1 Κίνητρο και σχεδιασμός λέξεων-κλειδιών . . . . . 77
5.2.2 Τροποποίηση του συνόλου δεδομένων με λέξεις-κλειδιά . . . . . 77
5.2.3 Χρήση ενισχυμένου μοντέλου στην τελική εφαρμογή . . . . . 78
5.3 Αξιολόγηση Βελτιωμένου Μοντέλου . . . . . 78
5.3.1 Εκπαίδευση και δοκιμή του Flash.75.8.1-k . . . . . 78
5.3.2 Βαθμολόγηση των απαντήσεων του μοντέλου . . . . . 79
5.4 Ανάπτυξη και αξιολόγηση πραγματικούς χρήστες . . . . . 81
5.4.1 Ενσωμάτωση του chatbot στην πλατφόρμα *ENERGATE* . . . . . 81
5.4.2 Ανθρώπινη αξιολόγηση της απόδοσης του βοηθού . . . . . 81
5.4.3 Σύνοψη Αποτελεσμάτων . . . . . 82

**III Επίλογος** **85**

**6 Επίλογος** **87**
6.1 Ακαδημαϊκή συμβολή . . . . . 87
6.2 Περιορισμοί και μελλοντική εργασία . . . . . 88
6.3 Τελικές παρατηρήσεις . . . . . 88

**Παραρτήματα** **91**

**Α΄ Συμπληρωματικές Τεχνικές Έννοιες** **93**
Α.1 Βελτιστοποίηση και Γενίκευση . . . . . 93
Α.1.1 Δεεπ Λεαρνινγ και η εποχή των τρανσφορμερς . . . . . 94
Α.2 Επεξεργασία Φυσικής Γλώσσας . . . . . 94
Α.2.1 Η Διαδικασία της Επεξεργασίας Φυσικής Γλώσσας . . . . . 94
Α.2.2 Σημασιολογία Κατανομής (Distributional Semantics) . . . . . 96
Α.3 Αρχιτεκτονική Transformer . . . . . 97
Α.3.1 Προσοχή με κλικακώμενο Εσωτερικό Γινόμενο (Scaled Dot-Product Attention) . . . . . 97
Α.3.2 Positional Encoding . . . . . 98
Α.3.3 Encoder-decoder stack . . . . . 98
Α.3.4 Παραλλαγές και ερμηνείες . . . . . 99

**Β΄ Δοκιμή τοπικής χρήσης LLaMA** **101**

**Γ΄ Python SDK για Vertex AI Tuning** **103**

**Δ΄ Αντιστοίχιση λέξεων-κλειδιών με προτροπές** **107**

**Βιβλιογραφία** **119**

**Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια** **121**

# Κατάλογος Σχημάτων


1.1 Ροή εργασίας της υλοποίησης και αξιολόγησης της διπλωματικής εργασίας . 24

2.1 Κύκλος ζωής LLM: επιμέλεια δεδομένων, προ-εκπαίδευση, ευθυγράμμιση (SFT + RLHF), προσαρμογή τομέα και ανάπτυξη με ονλινε συμπερασμό. Σχήμα 3 από Bommasani κ.α. [1]. . . . . 30

4.1 Ροή εργασίας υλοποίησης για την εκπαίδευση και την αξιολόγηση μοντέλων . 57
4.2 Βήμα 1 Fine-tuning Εργασίας: Επιλογή ονόματος fine-tuned μοντέλου, βασικού μοντέλου και περιοχής. Εάν οι παραμέτροι δεν λάβουν τιμή, τότε αποκτούν την προκαθορισμένη τιμή. . . . . 68
4.3 Βήμα 2 Fine-tuning Εργασίας: Επιλογή συνόλου δεδομένων εκπαίδευσης και (προαιρετικά) επικύρωσης. . . . . 69
4.4 Μετρικές κατά τη διάρκεια μίας tuning εργασίας. . . . . 69
4.5 Μετρικές στο τέλος μίας tuning εργασίας. . . . . 70

5.1 Κατανομή Βαθμολόγησης 60 ερωτημάτων (διακριτές κατηγορίες) . . . . . 80
5.2 Κατανομή Βαθμολόγησης 60 ερωτημάτων (ομαδοποιημένες κατηγορίες) . . . 80
5.3 Chatbot ενεργό στην πρώτη καρτέλα προσθήκης ενός κτηρίου, *Building Details* 82
5.4 Chatbot ενεργό στην καρτέλα τελικής προσφοράς, *Finalised Offer* . . . . . 82
5.5 Κατανομή Βαθμολόγησης σε 30 ερωτήματα πραγματικών χρηστών . . . . . 83

Α΄.1 Encoder–decoder transformer. Σχήμα 1 από Vaswani και συν. [2]. . . . . 98

# Κατάλογος Πινάκων

3.1 Συγκριτική επισκόπηση των LLM: Διαθεσιμότητα Πόρων, Ποιότητα Τεκμηρίωσης και Ευκολία Ανάπτυξης . . . . . 52
3.2 Συγκριτική επισκόπηση των LLM με βάση την απόδοση τους σε διάφορους δείκτες αναφοράς (IFEval, BBH, GPQA, MuSR, MMLU-PRO) . . . . . 55

4.1 Απαιτήσεις VRAM μοντέλων LlaMA για αποκλειστική εκτέλεση σε GPU με κβαντοποίηση 4-Bit . . . . . 59
4.2 Τιμές Απωλειών Επικύρωσης Άπληστης Αναζήτησης για Πλήθος Εποχών . . . 71
4.3 Τιμές Απωλειών Επικύρωσης Άπληστης Αναζήτησης για Μέγεθος Προσαρμογέα 71
4.4 Τιμές Απωλειών Επικύρωσης Άπληστης Αναζήτησης για Ρυθμό Μάθησης . . 71

5.1 Συγκριτικός Πίνακας μετρικών BLEU και ROUGE για τα μοντέλα `Flash.75.8.1` και `Flash.75.8.1-k` . . . . . 79

Κεφάλαιο 1

# Εισαγωγή

## 1.1 Η πρόκληση της ενεργειακής αναβάθμισης

Η κλιματική αλλαγή είναι από τα πλέον σημαντικά προβλήματα που αντιμετωπίζει η ανθρωπότητα τα τελευταία χρόνια. Πολλές χώρες έχουν αντιληφθεί τον κίνδυνο που σημαίνουν η αύξηση της θερμοκρασίας, τα ακραία καιρικά φαινόμενα και το φαινόμενο του θερμοκηπίου και έχουν προβεί σε ενέργειες καταπολέμησής τους. Συγκεκριμένα, η Ευρωπαϊκή Ένωση έχει θέσει αυστηρούς στόχους όσον αφορά τη βιωσιμότητα και το κλίμα, μέσω της ενεργειακής μετάβασης σε ανανεώσιμες πηγές ενέργειας. Η στρατηγική αυτή επικεντρώνεται στη μείωση των εκπομπών αερίων του θερμοκηπίου, μεταβιβάζοντας την εξάρτηση των διαφόρων υποδομών από τα ορυκτά καύσιμα σε ανανεώσιμες πηγές ενέργειας και γενικότερα βιώσιμων πρακτικών κατανάλωσης. Αυτή η αλλαγή στις υποδομές κάνει τα κτίρια πιο ενεργειακά αποδοτικά.

Η κλιματική αλλαγή έχει ενταθεί τον τελευταίο μισό αιώνα, με τις παγκόσμιες εκπομπές αερίων του θερμοκηπίου να αυξάνονται απότομα και να μην έχουν ακόμη κορυφωθεί [3]. Ως απάντηση, η ΕΕ υιοθέτησε ένα ολοκληρωμένο πλαίσιο πολιτικής για το κλίμα και την ενέργεια.

Η αναθεώρηση του 2023 των *Οδηγιών για την Ενεργειακή Αποδοτικότητα* (Energy Efficiency Directive) θέτει δεσμευτικό στόχο για πρόσθετη μείωση της τελικής κατανάλωσης ενέργειας κατά 11,7 % έως το 2030, ενώ η επανεξέταση του 2024 των *Οδηγιών για την Ενεργειακή Απόδοση Κτιρίων* (Energy Perfomance of Buildings Directive ή EPBD) απαιτεί μείωση κατά 60% των εκπομπών αερίων του θερμοκηπίου στον κτιριακό τομέα έως το 2030 και μηδενικές εκπομπές κτιρίων έως το 2050 [4, 5].

Οι πιο δημοφιλείς πρακτικές ενεργειακής μετάβασης (ή αναβάθμισης) κτηρίων περιλαμβάνουν προσθήκη καλύτερης μόνωσης, εγκατάσταση (ενεργειακά αποδοτικών) συστημάτων θέρμανσης και ψύξης (Heating, Ventilation and Air Conditioning ή HVAC), υιοθέτηση ανανεώσιμων πηγών ενέργειας (όπως φωτοβολταϊκά συστήματα) και προηγμένων τεχνολογιών διαχείρισης ενέργειας (Energy Management System ή EMS). Τα έργα ενεργειακής αναβάθμισης συγκαταλέγονται μεταξύ των πλέον αποδοτικών μεθόδων για την επίτευξη των στόχων σε επίπεδο ΕΕ. Προσφέρουν σε βάθος χρόνου οικονομικά οφέλη στους ιδιοκτήτες των υποδομών, ενώ ταυτόχρονα βελτιώνουν σε σημαντικό βαθμό τις περιβαλλοντικές επιπτώσεις. Οι μετα-αναλύσεις δείχνουν ότι τα μέτρα αυτά παρέχουν εσωτερικούς συντελεστές απόδοσης της τάξης του 5-25% για τα περισσότερα ευρωπαϊκά κλίματα και περιόδους απόσ-

βεσης κάτω των δέκα ετών όταν συνδυάζονται στρατηγικά [6, 7].

Για την επίτευξη αυτών των αναβαθμίσεων υπάρχουν πολλοί ενδιαφερόμενοι (stakeholders), όπως δημόσιοι ή ιδιωτικοί ιδιοκτήτες κτιρίων, εταιρείες ενεργειακών υπηρεσιών (Energy Perfomance COmpany ή ESCO), εργολάβοι (ή μηχανικοί) και επενδυτές. Η οικονομική υποστήριξη βασίζεται σε δάνεια, επιχορηγήσεις, Συμπράξεις Δημοσίου και Ιδιωτικού Τομέα (ΣΔΙΤ) και συμβάσεις ενεργειακής απόδοσης (Energy Perfomance Certificate ή EPC). Οι ενδιαφερόμενοι ποικίλουν δημογραφικά, σε άποψη ηλικίας, τεχνικής κατάρτισης και εξοικείωσης με ψηφιακές τεχνολογίες, ενώ κάθε ενδιαφερόμενος έχει μοναδικές ανάγκες, τεχνογνωσία και προσδοκίες. Αυτή η ποικιλομορφία οδηγεί συχνά σε εσφαλμένη επικοινωνία, καθυστερήσεις και αναποτελεσματικότητα στις φάσεις υλοποίησης του έργου.

Εκτός από την πληθώρα των διαφορετικών ενδιαφερομένων, μία άλλη πρόκληση για την υλοποίηση των έργων ενεργειακής αναβάθμισης αποτελεί η διαφοροποίηση των προτύπων, των κανονισμών και των διαδικασιών ενεργειακής απόδοσης στις διάφορες χώρες της ΕΕ. Η έλλειψη τυποποίησης και προδιαγραφών δυσχεραίνει τη συνεργασία ανάμεσα σε ενδιαφερόμενους διαφορετικών χωρών.

## 1.2 Ο ρόλος των ψηφιακών τεχνολογιών

Αυτά τα προβλήματα έρχονται να λύσουν οι ψηφιακές τεχνολογίες, οι οποίες αποκτούν ολοένα και μεγαλύτερη ανταπόκριση στον τομέα της ενεργειακής απόδοσης και αναβάθμισης. Οι διαδικτυακές πλατφόρμες, όπως η *ENERGATE* συμβάλλουν στην απλούστευση των διαδικασιών και στη βελτίωση της επικοινωνίας μεταξύ των ενδιαφερομένων.

Η *ENERGATE* είναι μια πλατφόρμα για επενδύσεις ενεργειακής αποδοτικότητας σε κτίρια, που αναπτύχθηκε με την υποστήριξη του προγράμματος LIFE[1] της ΕΕ [8]. Η πλατφόρμα συγκεντρώνει δεδομένα έργων και συνδέει απευθείας τους ιδιοκτήτες κτιρίων, τους υπεύθυνους υλοποίησης ανακαινίσεων και τους χρηματοδότες, απλοποιώντας τη σύναψη συμφωνιών και απελευθερώνοντας ιδιωτικά κεφάλαια για ριζικές ανακαινίσεις. Στόχος της είναι να αυξήσει το ποσοστό ανακαίνισης του κτιριακού αποθέματος της Ευρώπης, σύμφωνα με την αρχή «Efficiency First» της Ένωσης και τους κλιματικούς στόχους για το 2030 [9].

Ωστόσο, πολλοί ενδιαφερόμενοι θεωρούν ότι αυτές οι πλατφόρμες είναι δύσκολες να πλοηγηθούν, με αποτέλεσμα να χρειάζονται υποστήριξη από ανθρώπινο προσωπικό, γεγονός που μειώνει την αποτελεσματικότητά τους. Παρά τις υποσχέσεις τους, αυτές οι πλατφόρμες συχνά απογοητεύουν τους χρήστες λόγω (i) της πολυγλωσσικής ορολογίας, (ii) της εξειδικευμένης τεχνικής ή χρηματοοικονομικής ορολογίας και (iii) των ειδικών ρυθμιστικών μέτρων κάθε χώρας [10].

Πρόσφατες μελέτες υπογραμμίζουν τη σημασία της ενσωμάτωσης των εικονικών βοηθών σε σύνθετες πλατφόρμες για τη βελτίωση της εμπειρίας του χρήστη. Οι Galanakis κ.α. τονίζουν ότι οι εικονικοί βοηθοί σε περιβάλλοντα Βιομηχανίας 4.0[2] παρέχουν τόσο Εικονική

[1] Μτφ. από Γαλλικά, *L'Instrument Financier pour L'Enviroment*: Programme for the Environment and Climate Action ή Πρόγραμμα για το Περιβάλλον και τη Δράση για το Κλίμα.

[2] Η Βιομηχανία 4.0 (Industry 4.0), γνωστή και ως Τέταρτη Βιομηχανική Επανάσταση (Fourth Industrial Revolution ή 4IR), είναι η επόμενη φάση της ψηφιοποίησης του ψηφιοποίηση του τομέα της βιομηχανικής παραγωγής, η οποία καθοδηγείται από ρηξικέλευθες τάσεις, όπως η άνοδος των δεδομένων και της συνδεσιμότητας, η ανάλυση δεδομένων, η αλληλεπίδραση ανθρώπου-μηχανής και οι βελτιώσεις στη ρομποτική [11].

Βοήθεια — προσφέροντας πληροφορίες σε πραγματικό χρόνο και με βάση το πλαίσιο — όσο και Φυσική Βοήθεια — υποστηρίζοντας την εκτέλεση εργασιών. Αυτές οι διπλές λειτουργίες μπορούν να συμβάλουν καθοριστικά στην αντιμετώπιση των πολύπλευρων προκλήσεων που αντιμετωπίζουν οι χρήστες σε πλατφόρμες όπως η *ENERGATE* [12].

Η αντιμετώπιση αυτών των ζητημάτων απαιτεί καινοτόμες ψηφιακές λύσεις που να ανταποκρίνονται στις διαφορετικές ανάγκες όλων των ενδιαφερομένων. Οι πρόσφατες τεχνολογικές εξελίξεις, ιδίως στον τομέα της τεχνητής νοημοσύνης (Artificial Intelligence ή AI) και της επεξεργασίας φυσικής γλώσσας (Natural Language Processing ή NLP), έχουν οδηγήσει σε εξελιγμένους εικονικούς βοηθούς (Virtual Agents ή VA) και τεχνολογίες chatbot, οι οποίες μπορούν να συμβάλουν στη γεφύρωση του χάσματος μεταξύ των ενδιαφερομένων και της ψηφιοποίησης του τομέα της ενέργειας, διευκολύνοντας έτσι την πρόσβαση και χρήση νέων τεχνολογιών.

Ως εικονικός βοηθός (ΕΒ) μπορεί να οριστεί ένας πράκτορας λογισμικού που εκτελεί ενέργειες ή υπηρεσίες βάσει εντολών ή ερωτήσεων, ο οποίος μπορεί να ενεργοποιείται με φωνητικές εντολές ή εισόδους βάσει κειμένου. Οι Galanakis και συν. αποδεικνύουν πως οι ΕΒ ενσωματώνουν τεχνολογίες όπως η μηχανική μάθηση (Machine Learning ή ML), η αναγνώριση ομιλίας, η απάντηση ερωτήσεων, η διαχείριση διαλόγων, η παραγωγή γλώσσας, η σύνθεση κειμένου σε ομιλία, η εξόρυξη δεδομένων, η ανάλυση, η εξαγωγή συμπερασμάτων και η προσαρμογή [12]. Οι τεχνολογίες αυτές βασίζονται κυρίως στα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models ή LLMs), όπως LLaMA, BLOOM, GPT ή Gemini [13, 14, 15, 16].

Πολύ παρόμοια με τους ΕΒ είναι τα chatbots, προγράμματα υπολογιστών σχεδιασμένα να απαντούν σε ερωτήσεις και να συνομιλούν με ανθρώπους μέσω κειμένου ή φωνής. Καθώς ένας ΕΒ λειτουργεί ως διαμεσολαβητής μεταξύ του χρήστη και μιας συγκεκριμένης υπηρεσίας ή εφαρμογής για την εκπλήρωση της πρόθεσης του χρήστη και τα chatbots είναι ένα μέσο αλληλεπίδρασης μεταξύ ανθρώπων και υπολογιστών, ένα chatbot μπορεί να θεωρηθεί ως ένα θεμελιώδες κομμάτι ενός ΕΒ που επιτρέπει να δοθούν οδηγίες στον βοηθό με φυσικό τρόπο. Στο πλαίσιο της διαχείρισης ανθρώπινου δυναμικού, οι Kothapalli και συν. αποδεικνύουν ότι τα chatbots βελτιστοποιούν αποτελεσματικά διαδικασίες όπως η ένταξη και η υποστήριξη των εργαζομένων. Τα ευρήματά τους υποδηλώνουν ότι τέτοιες εφαρμογές μπορούν να προσαρμοστούν για να βοηθήσουν χρήστες με διαφορετικά επίπεδα ψηφιακής παιδείας, ένα σημαντικό στοιχείο για τη διαφορετική βάση χρηστών της *ENERGATE* [17].

Οι ΕΒ παρέχουν διαισθητικές και φιλικές προς τον χρήστη λύσεις, ακόμη και για ενδιαφερόμενους που είναι λιγότερο εξοικειωμένοι με την ψηφιακή τεχνολογία [12]. Μπορούν να χειριστούν αποτελεσματικά τα εμπόδια επικοινωνίας μεταφράζοντας πολύπλοκες τεχνικές πληροφορίες σε απλή, σαφή και κατανοητή γλώσσα. Αυτό διευκολύνει την αλληλεπίδραση των ενδιαφερομένων με διαφορετικά επίπεδα εμπειρογνωμοσύνης με την πλατφόρμα και μειώνει τις παρανοήσεις.

Η έρευνα των Chaves και Gerosa δείχνει ότι η ενσωμάτωση ανθρωπομορφικών χαρακτηριστικών στα chatbots, όπως απαντήσεις που μοιάζουν με ανθρώπινες ή συμπεριφορές που επιζητούν διευκρινίσεις, μπορεί να ενισχύσει σημαντικά την εμπιστοσύνη των χρηστών και τα ποσοστά υιοθέτησης. Αυτή η προσέγγιση μπορεί να είναι ιδιαίτερα επωφελής για χρήστες που δεν είναι εξοικειωμένοι με τις ψηφιακές τεχνολογίες, καθώς προάγει μια πιο

οικεία και ελκυστική αλληλεπίδραση [18].

Παρά τα πολυάριθμα πλεονεκτήματά τους, οι ΕΒ έχουν ορισμένους περιορισμούς. Στις συνηθέστερες προκλήσεις συγκαταλέγονται η δυσκολία στην ενσωμάτωσή τους στις ψηφιακές πλατφόρμες, η ακριβής κατανόηση σύνθετων ή εξειδικευμένων ερωτημάτων και η διατήρηση της ασφάλειας των δεδομένων των χρηστών. Επιπλέον, οι ΕΒ για να είναι αποτελεσματικοί πρέπει να λειτουργούν σε πραγματικό χρόνο, δηλαδή να απαντούν γρήγορα και με ακρίβεια στα αιτήματα των χρηστών [19].

Οι Adam κ.α. κατηγοριοποιούν τα chatbots υπηρεσιών σε ενημερωτικά, συναλλακτικά και συμβουλευτικά, σημειώνοντας ότι ο αρθρωτός (modular) σχεδιασμός προσαρμοσμένος σε συγκεκριμένες εργασίες ενισχύει την αποτελεσματικότητα. Η εφαρμογή αυτού του πλαισίου στην *ENERGATE* θα μπορούσε να περιλαμβάνει την ανάπτυξη εξειδικευμένων ενοτήτων για την υποστήριξη διακριτών αναγκών των χρηστών, όπως η πλοήγηση σε κανονιστικές πληροφορίες ή ο οικονομικός σχεδιασμός [20].

Για να μπορέσουν οι ΕΒ να αξιοποιήσουν πλήρως τις δυνατότητές τους, απαιτούνται συνεχείς βελτιώσεις. Η βιβλιογραφία δεν είναι εκτενής, καθώς είναι σχετικά καινούργιος τομέας, αλλά οι υπάρχουσες ερευνητικές εργασίες δείχνουν πως εξελίξεις στα αρχιτεκτονικά στοιχεία των βοηθών όπως η κατανόηση φυσικής γλώσσας (Natural Language Understanding ή NLU), τα συστήματα διαχείρισης διαλόγου (dialog management systems) και η παραγωγή φυσικής γλώσσας (Natural Language Generation ή NLG) μπορούν να υποστηρίξουν πιο φυσιολογικές αλληλεπιδράσεις. Η αξιολόγηση αυτών των συστημάτων με κατάλληλες μετρικές -όπως η ακρίβεια, το F1-score [21], το BLEU [22] και η ικανοποίηση των χρηστών- θα είναι απαραίτητη για την εξασφάλιση αξιόπιστων επιδόσεων.

Η εξέταση των AI chatbots στις χρηματοοικονομικές υπηρεσίες από τους Sharma και Prasad αποκαλύπτει ότι η επιτυχία εξαρτάται από την ακριβή κατανόηση της φυσικής γλώσσας και την ανταπόκριση σε πραγματικό χρόνο. Αυτές οι πληροφορίες υπογραμμίζουν τη σημασία της προσαρμογής (finetuning) των ΕΒ για την αποτελεσματική διαχείριση ερωτήσεων που αφορούν συγκεκριμένους τομείς, διασφαλίζοντας ότι ανταποκρίνονται στις λεπτές απαιτήσεις πλατφορμών όπως η *ENERGATE* [23]. Τέλος, η προσθήκη ανθρωπομορφικών χαρακτηριστικών, όπως η συμπεριφορά αναζήτησης διευκρινίσεων, μπορεί να ενισχύσει την εμπιστοσύνη και την υιοθέτησή τους από τους χρήστες, ιδίως για εκείνους με υψηλή ανάγκη για ανθρώπινη αλληλεπίδραση (Need For Human Interaction ή NFHI), κάνοντάς τον βοηθό πιο ελκυστικό σε άτομα που δεν είναι εξοικειωμένα με την ψηφιακή τεχνολογία [18, 20, 23].

## 1.3 Η πλατφόρμα *ENERGATE* και ο στόχος της διπλωματικής εργασίας

Μία από τις διαδικτυακές πλατφόρμες που διευκολύνουν την εφαρμογή της ενεργειακής αναβάθμισης και η οποία συνεργάζεται με την ΕΕ είναι η *ENERGATE*. Πρόκειται για ένα ψηφιακό εργαλείο που έχει σχεδιαστεί για την υποστήριξη επενδύσεων σε έργα ενεργειακής αποδοτικότητας κτηρίων. Παρέχει στους χρήστες υπηρεσίες που σχετίζονται με την αξιολόγηση έργων, τη χρηματοδότηση, τη διαχείριση δεδομένων και τη συμμόρφωση με τους

ενεργειακούς κανονισμούς. Ενώ η πλατφόρμα είναι πλούσια σε χαρακτηριστικά, μπορεί να θεωρηθεί πολύπλοκη και συντριπτική από τους χρήστες (ιδιοκτήτες κτηρίων, μηχανικοί, επενδυτές και δημόσιες αρχές) με περιορισμένες τεχνικές γνώσεις ή μη εξοικείωση με τις ψηφιακές διεπαφές.

Για την αντιμετώπιση αυτών των προκλήσεων, η παρούσα διπλωματική εργασία προτείνει το σχεδιασμό και την ανάπτυξη ενός ΕΒ που βασίζεται σε ένα μεγάλο γλωσσικό μοντέλο. Ο βοηθός θα ενσωματωθεί στην πλατφόρμα *ENERGATE* και θα χρησιμεύσει ως οδηγός για να βοηθήσει τους χρήστες να πλοηγηθούν αποτελεσματικότερα στις λειτουργίες της. Θα παρέχει βήμα προς βήμα οδηγίες, θα αποσαφηνίζει άγνωστους όρους, θα απαντά σε ερωτήσεις των χρηστών και θα βοηθά στην ολοκλήρωση πολύπλοκων ροών εργασίας.

Στόχος είναι η δημιουργία ενός εργαλείου που θα βελτιώνει την εμπειρία των χρηστών, θα μειώνει την ανάγκη για ανθρώπινη υποστήριξη και θα αυξάνει την προσβασιμότητα της πλατφόρμας για χρήστες με διαφορετικά επίπεδα εμπειρίας. Με την πιλοτική δοκιμή του βοηθού με πραγματικούς χρήστες, η παρούσα εργασία συλλέγει ανατροφοδότηση και αξιολογεί την αποτελεσματικότητά του στην επίλυση προβλημάτων ευχρηστίας.

Η παρούσα διπλωματική εργασία διερευνά επίσης την απόδοση διαφορετικών LLM (όπως LLaMA και BLOOM), πειραματίζεται με δομές γνώσης και αξιολογεί διάφορες στρατηγικές για τη λεπτομερή ρύθμιση (fine-tuning) του βοηθού. Ιδιαίτερη έμφαση δίνεται στην αξιολόγηση διαφορετικών δομών δεδομένων εκπαίδευσης και πώς αυτές μπορούν να οδηγήσουν σε βελτιωμένα αποτελέσματα, με σκοπό να λαμβάνονται υπόψη η περιήγηση του χρήστη εντός της πλατφόρμας σε πραγματικό χρόνο. Το τελικό αποτέλεσμα είναι ένα λειτουργικό πρωτότυπο που θα συμβάλει στο να γίνει η πλατφόρμα *ENERGATE* πιο φιλική προς το χρήστη και πιο αποτελεσματική.

## 1.4 Οργάνωση του τόμου

Η διπλωματική εργασία αυτή είναι οργανωμένη σε έξι κεφάλαια, καθένα από τα οποία ασχολείται με ένα ξεχωριστό στοιχείο της έρευνας σχετικά με το σχεδιασμό, τη βελτιστοποίηση (fine-tuning) και την ανάπτυξη ενός εικονικού βοηθού βασισμένου σε ένα μεγάλο γλωσσικό μοντέλο (LLM) για την πλατφόρμα ενεργειακής απόδοσης *ENERGATE*.

Το Κεφάλαιο 2 παρέχει το θεωρητικό υπόβαθρο που είναι απαραίτητο για την κατανόηση των τεχνολογιών που χρησιμοποιούνται σε αυτή την εργασία. Εξηγεί τις βασικές έννοιες της τεχνητής νοημοσύνης και της μηχανικής μάθησης, περιγράφει τη δομή και τις διαδικασίες εκπαίδευσης των μεγάλων γλωσσικών μοντέλων και εισάγει μετρήσεις αξιολόγησης και ηθικές παραμέτρους που σχετίζονται με την ανάπτυξη τους σε ρυθμιζόμενα ψηφιακά περιβάλλοντα.

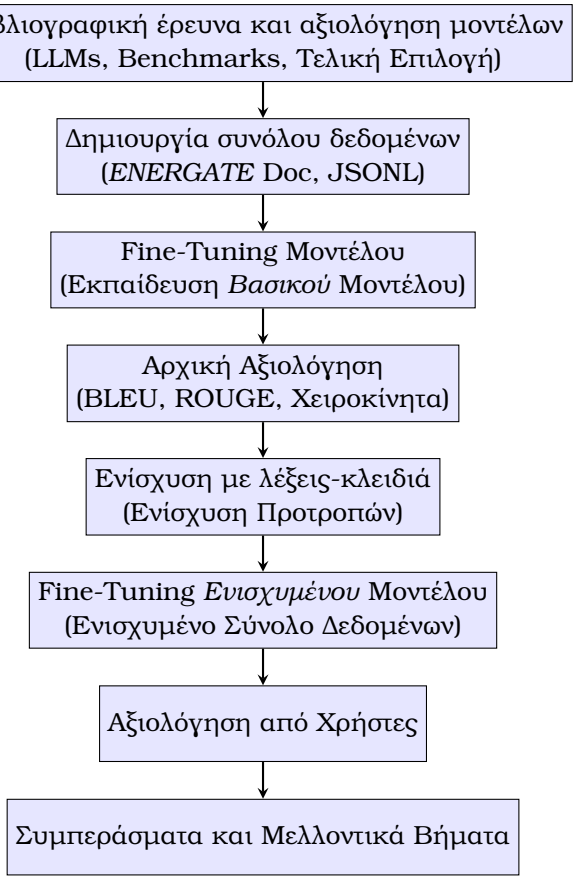
Το Κεφάλαιο 3 περιγράφει το μεθοδολογικό πλαίσιο που υιοθετήθηκε για την επιλογή ενός κατάλληλου βασικού μοντέλου για τη βελτιστοποίηση. Παρουσιάζει μια συγκριτική ανάλυση των διαθέσιμων LLMs με βάση ποιοτικά χαρακτηριστικά (όπως τεκμηρίωση και ευκολία εφαρμογής) και ποσοτικά κριτήρια απόδοσης, που τελικά οδήγησαν στην επιλογή του Gemini για αυτό το έργο.

Το Κεφάλαιο 4 τεκμηριώνει τη διαδικασία υλοποίησης. Ξεκινά με μια ανεπιτυχή προσπάθεια χρήσης του LLaMA σε τοπικό επίπεδο και συνεχίζει με την περιγραφή της επιτυχημένης βελτιστοποίησης ενός μοντέλου Gemini στο Google Cloud Vertex AI. Περιγράφει λεπτομερώς

την προετοιμασία των δεδομένων εκπαίδευσης, τη στρατηγική επαύξησης, τη ρύθμιση των υπερπαραμέτρων και τις διαδικασίες αξιολόγησης, συμπεριλαμβανομένης της χρήσης προτροπών ενισχυμένων με λέξεις-κλειδιά για την προσομοίωση του περιβάλλοντος του χρήστη εντός της πλατφόρμας.

Το Κεφάλαιο 5 παρουσιάζει τα αποτελέσματα της αξιολόγησης του μοντέλου. Περιλαμβάνει αυτοματοποιημένες μετρήσεις απόδοσης (BLEU, ROUGE), χειροκίνητες αξιολογήσεις ποιότητας απόκρισης και πραγματικά σχόλια χρηστών από ένα πρωτότυπο ενσωματωμένο στην *ENERGATE*. Τα αποτελέσματα καταδεικνύουν την πρακτική αποτελεσματικότητα του βοηθού και υπογραμμίζουν τις βελτιώσεις που επιτεύχθηκαν μέσω του σχεδιασμού προτροπών που λαμβάνουν υπόψη το περιβάλλον.

Τέλος, το Κεφάλαιο 6 συνοψίζει τις κύριες συνεισφορές της εργασίας, αναλύει τις επιπτώσεις των LLM εικονικών βοηθών στις πλατφόρμες ψηφιοποίησης της ενέργειας και σκιαγραφεί τις κατευθύνσεις για μελλοντική εργασία, όπως η επέκταση των πολυγλωσσικών δυνατοτήτων του βοηθού και η βελτίωση των μηχανισμών συμμόρφωσης με τους κανονισμούς.




Σχήμα 1.1: *Ροή εργασίας της υλοποίησης και αξιολόγησης της διπλωματικής εργασίας*

# Μέρος I

# Θεωρητικό Μέρος

Κεφάλαιο 2

# Θεωρητικό υπόβαθρο

Αυτό το κεφάλαιο θέτει τις εννοιολογικές και τεχνικές βάσεις που είναι απαραίτητες για την κατανόηση και την κατασκευή ενός εικονικού βοηθού που λειτουργεί με μεγάλα γλωσσικά μοντέλα. Ξεκινά με την περιγραφή των βασικών αρχών της τεχνητής νοημοσύνης και της μηχανικής μάθησης, συμπεριλαμβανομένων των κύριων παραδειγμάτων μάθησης που υποστηρίζουν την ανάπτυξη μοντέλων. Στη συνέχεια, επικεντρώνεται στη δομή, την εκπαίδευση και την εξειδίκευση των μεγάλων γλωσσικών μοντέλων, επισημαίνοντας τη διαδικασία από την προ-εκπαίδευση έως την τελειοποίηση και την ανάπτυξη.

Στις επόμενες ενότητες περιγράφεται λεπτομερώς ο τρόπος μέτρησης της απόδοσης των μοντέλων μέσω αυτομάτων μετρήσεων και δοκιμών αναφοράς (benchmarks), και ολοκληρώνεται με την εξέταση των ηθικών, νομικών και κοινωνικών παραμέτρων που σχετίζονται με την ανάπτυξη τέτοιων συστημάτων σε ρυθμιζόμενα περιβάλλοντα, όπως οι πλατφόρμες ενεργειακής απόδοσης. Για μια πιο λεπτομερή περιγραφή των υποκείμενων αλγορίθμων και τεχνικών εφαρμογών, βλ. Παράρτημα Α΄ Συμπληρωματικές Τεχνικές Έννοιες.

## 2.1 Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

Η Τεχνητή Νοημοσύνη (ΤΝ) είναι ο ερευνητικός τομέας που μελετά τον τρόπο κατασκευής μηχανών ικανών να εκτελούν εργασίες που απαιτούν νοημοσύνη όταν εκτελούνται από ανθρώπους [24]. Ο όρος αυτός επινοήθηκε στο Dartmouth Summer Research Project του 1956, ένα γεγονός που αναγνωρίζεται ευρέως ως η επίσημη γέννηση του κλάδου [25].

Το πρώτο πρόγραμμα μάθησης (που παρουσιάστηκε σε πειραματική εφαρμογή σε mainframe της IBM) ήταν το πρόγραμμα παίκτη του παιχνιδιού ντάμας (checkers) του Arthur Samuel, το οποίο απέδειξε ότι ένας υπολογιστής μπορεί να βελτιώσει την απόδοσή του μέσω της εμπειρίας και όχι μόνο μέσω ρητού προγραμματισμού [26]. Το έργο του Samuel εισήγαγε τον όρο «Μηχανική Μάθηση» (ΜΜ) και μετατόπισε την προσοχή από τους κανόνες μάθησης μέσω προγραμματισμού στην βελτιστοποίηση βάσει δεδομένων. Ο Tom Mitchell έδωσε αργότερα τον κανονικό ορισμό του πεδίου: μάθηση από την εμπειρία $E$ (experience) σε εργασίες $T$ (tasks) όπως μετράται από την απόδοση $P$ (performance), θεμελιώνοντας τις σύγχρονες μεθόδους έρευνας [27].

Η σύγχρονη τεχνητή νοημοσύνη περιλαμβάνει, μεταξύ άλλων, την αντίληψη, τη συλλογιστική, τον σχεδιασμό, τη γλώσσα, τη ρομποτική και τη μάθηση [24]. Η μηχανική μάθηση κυριαρχεί πλέον σε πρακτικές εφαρμογές, επιτρέποντας την αυτόματη εξαγωγή

χρήσιμων αναπαραστάσεων από ακατέργαστα δεδομένα [28]. Οι κλασικές τεχνικές τεχνητής νοημοσύνης, όπως η συμβολική συλλογιστική (symbolic reasoning)[1], παραμένουν σχετικές, αλλά συνδυάζονται όλο και περισσότερο με συστήματα μάθησης για να αξιοποιήσουν συμπληρωματικά πλεονεκτήματα.

### 2.1.1 Παραδείγματα μάθησης

Η μηχανική μάθηση διακρίνει τέσσερα βασικά παραδείγματα μάθησης, καθένα από τα οποία χαρακτηρίζεται από τη διαθεσιμότητα επισημασμένης ανατροφοδότησης (labelled feedback) και τη φύση της αλληλεπίδρασης με το περιβάλλον [30].

- Η *εποπτευόμενη μάθηση* (supervised learning) μαθαίνει μια αντιστοίχιση (mapping) $f : \mathcal{X} \to \mathcal{Y}$ ελαχιστοποιώντας την απώλεια μεταξύ των αποτελεσμάτων του μοντέλου (model outputs) και των επισημανμένων στόχων (labelled targets). Η ταξινόμηση εικόνων και η αυτόματη μετάφραση είναι τυπικά παραδείγματα εποπτευόμενης μάθησης [31].

- Η *μη-εποπτευόμενη μάθηση* ανακαλύπτει υποκείμενες δομές (latent regularities) σε μη επισημασμένα δεδομένα, π.χ. ομαδοποίηση με $k$-means [31, 32], αποδίδοντας αναπαραστάσεις χρήσιμες για μεταγενέστερες εργασίες[2] [31].

- Η *αυτοεποπτευόμενη μάθηση* (self-supervised learning) μετατρέπει τμήματα της ακατέργαστης εισόδου σε ψευδοετικέτες επιτρέποντας τη χρήση τεράστιων μη σχολιασμένων συνόλων δεδομένων. Η πρόβλεψη κρυμμένων συμβόλων (masked-token prediction) στο BERT είναι ένα χαρακτηριστικό παράδειγμα [34].

- Η *ενισχυτική μάθηση* (reinforcement learning) μοντελοποιεί τη λήψη αποφάσεων ως μεγιστοποίηση της αναμενόμενης σωρευτικής ανταμοιβής σε μια διαδικασία λήψης αποφάσεων Markov, η οποία επιλύεται με μεθόδους κλίσης πολιτικής (policy-gradient) ή χρονικής διαφοράς (temporal-difference) [35].

## 2.2 Large Language Models

Τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models ή LLMs) είναι δίκτυα μετασχηματιστών (transformer networks) των οποίων ο αριθμός παραμέτρων υπερβαίνει περίπου το ένα δισεκατομμύριο, επιτρέποντας την εμφάνιση νέων ικανοτήτων, όπως η in-context μάθηση και η zero-shot γενίκευση[3] [1]. Οι νόμοι κλιμάκωσης (scaling laws) δείχνουν ότι η

[1] Η *συμβολική συλλογιστική* (symbolic reasoning), γνωστή και ως συμβολική τεχνητή νοημοσύνη ή κλασική τεχνητή νοημοσύνη, αναπαριστά τη γνώση ως ρητά σύμβολα και τα χειρίζεται μέσω της τυπικής λογικής ή της συμπερασματολογίας βάσει κανόνων. Παρέχει γρήγορη και επαληθεύσιμη συμπερασματολογία πάνω σε δομημένη γνώση [29].

[2] Μια *μεταγενέστερη* ή *μεταεκπαιδευτική* εργασία (downstream task) είναι μια εργασία στην οποία εφαρμόζεται ένα μοντέλο μετά την προεκπαίδευση. Συνήθως περιλαμβάνει το finetuning του προ-εκπαιδευμένου μοντέλου σε ένα μικρότερο, συγκεκριμένο σύνολο δεδομένων. Τέτοιες εργασίες περιλαμβάνουν: ταξινόμηση κειμένου (π.χ. ανίχνευση spam, ανάλυση συναισθήματος), αναγνώριση ονομαστικών οντοτήτων (Name Entity Recognition ή NER), απάντηση ερωτήσεων, μετάφραση ή περίληψη κειμένου και διάλογος (όπως στην περίπτωσή ενός εικονικού πράκτορα) [33].

[3] Η *μάθηση στο πλαίσιο* (in-context) αναφέρεται στην ικανότητα ενός μοντέλου να εκτελεί εργασίες απλά με βάση παραδείγματα που παρέχονται στην είσοδο χωρίς ενημέρωση παραμέτρων. Η *zero-shot γενίκευση* είναι

απώλεια δοκιμής (test loss) ακολουθεί μια φθίνουσα εξάρτηση (power-law) από το μέγεθος του μοντέλου, το μέγεθος του συνόλου δεδομένων και την υπολογιστική ισχύ. Ως εκ τούτου, οι βελτιώσεις στην απόδοση ακολουθούν την αύξηση των παραμέτρων από 1,5 δισεκατομμύρια του GPT-2 σε 1,56 τρισεκατομμύρια του Gemini 1.5-Ultra [37, 38, 36, 15, 39]. Η ενότητα αυτή τυποποιεί αρχικά τους στόχους της προ-εκπαίδευσης (pre-training) που προσδίδουν στα LLM γενική γλωσσική ικανότητα και, στη συνέχεια, εξετάζει τεχνικές ευθυγράμμισης που εξειδικεύουν αυτή την ικανότητα για ασφαλή, εξειδικευμένη ως προς το πεδίο αλληλεπίδραση.

### 2.2.1 Στόχοι προ-εκπαίδευσης

#### Μοντελοποίηση κρυμμένης γλώσσας

Το BERT και οι διάδοχοί του κρύβουν τυχαία το 15% των εισερχόμενων token και εκπαιδεύουν το μοντέλο να προβλέπει τα κομμάτια που λείπουν με βάση το αμφίδρομο πλαίσιο (bidirectional context) [34]. Η μοντελοποίηση κρυμμένης γλώσσας (Masked-Language Modelling ή MLM) αποδίδει ισχυρή κατανόηση σε επίπεδο πρότασης, αλλά απαιτεί ξεχωριστό αποκωδικοποιητή για τη δημιουργία/παραγωγή (generation), περιορίζοντας τους συνομιλητικούς πράκτορες (conversational agents).

#### Μοντελοποίηση αιτιώδους (αυτοπαλινδρομικής) γλώσσας

Τα μοντέλα τύπου GPT μεγιστοποιούν την πιθανότητα κάθε token με βάση όλα τα προηγούμενα tokens:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{n} \log p_\theta(x_t \mid x_{<t}).$$

Όπου $n$ είναι το μήκος της ακολουθίας, $x_t$ το token στη θέση $t$, $x_{<t}$ το πρόθεμα $(x_1, \ldots, x_{t-1})$ και $p_\theta$ η κατανομή του μοντέλου με παραμέτρους $\theta$. Ελαχιστοποιώντας το $\mathcal{L}_{\text{CLM}}$, το μοντέλο εκπαιδεύεται να προβλέπει το επόμενο token με βάση όλα τα προηγούμενα tokens. Το μονοκατευθυντικό πλαίσιο (unidirectional context) υποστηρίζει την ελεύθερη μορφή δημιουργίας που απαιτείται για τον διάλογο [36]. Οι instruction-tuned παραλλαγές (FLAN-T5) συνδυάζουν και τους δύο στόχους μέσω μικτών διαστημάτων αποθορυβοποίησης (mixed denoising spans) και αυτοπαλινδρομικών κεφαλών (autoregressive heads) [40].

#### Καθεστώς (regime) δεδομένων

Τα LLM απορροφούν τρισεκατομμύρια tokens που καλύπτουν web crawls[4], ακαδημαϊκά σώματα κειμένων, κώδικα και πολυγλωσσικά δεδομένα. Η απομάκρυνση διπλών δεδομένων[5] και το φιλτράρισμα τοξικότητας εξισορροπούν την κλίμακα με την ποιότητα [42]. Έγγραφα ειδικά για την *ENERGATE* μπορούν να προστεθούν ως μια ελαφριά φάση συνεχούς προεκπαίδευσης για την εισαγωγή γνώσης του τομέα χωρίς πλήρη επανεκπαίδευση.

η ικανότητα εκτέλεσης μιας νέας εργασίας χωρίς να έχουν δει τα δεδομένα με ετικέτες κατά τη διάρκεια της εκπαίδευσης, βασιζόμενη αποκλειστικά σε οδηγίες ή προτροπές σε φυσική γλώσσα [36].

[4] Αυτοματοποιημένα *spiders* που κατεβάζουν συστηματικά δημόσιες ιστοσελίδες για να δημιουργήσουν μεγάλης κλίμακας σώματα κειμένων, όπως το Common Crawl.

[5] Η *απομάκρυνση διπλών δεδομένων* ή *deduplication* αναφέρεται στο hash-based ή στο locality-sensitive φιλτράρισμα που αφαιρεί σχεδόν-διπλά έγγραφα για να μειώσει την απομνημόνευση (memorisation) και την μεροληψία (bias) [41].

### 2.2.2 Τεχνικές ευθυγράμμισης

Η προ-εκπαίδευση παρέχει ευρεία ικανότητα, αλλά όχι υπακοή στις εργασίες ή ασφάλεια. Οι μέθοδοι ευθυγράμμισης (alignment methods) καλύπτουν αυτό το κενό.

**Fine-tuning οδηγιών**

Η εποπτευόμενη λεπτομερής ρύθμιση (Supervised Fine-Tuning ή SFT) σε ζεύγη (<*οδηγία*, *απάντηση*>) διδάσκει στο μοντέλο να ακολουθεί εντολές φυσικής γλώσσας [43]. Τα σύνολα δεδομένων ειδικού τομέα (π.χ. προτροπές ροής εργασίας *ENERGATE*) προσαρμόζουν περαιτέρω τη συμπεριφορά του βοηθού.

**Ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση**

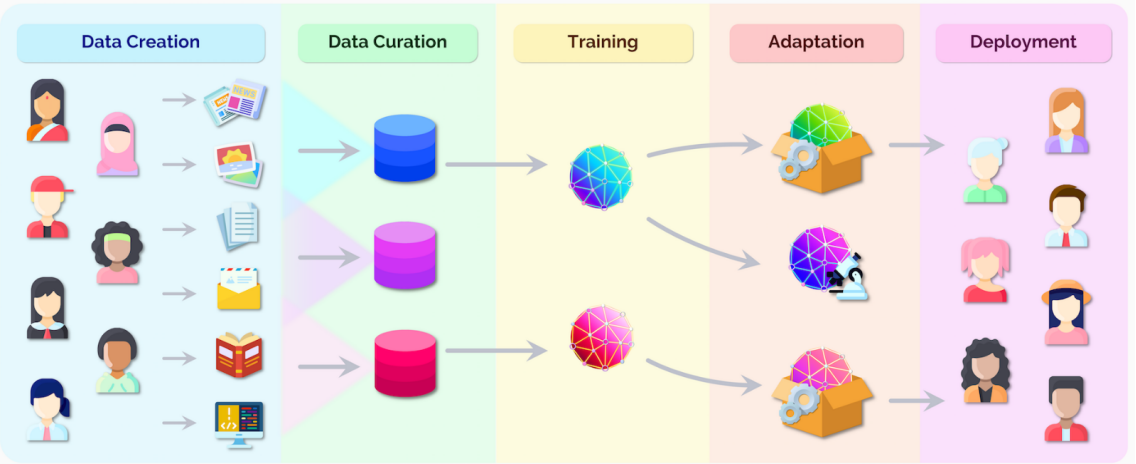
Το μοντέλο παράγει πρώτα υποψήφιες απαντήσεις, ένα μοντέλο ανταμοιβής (reward model) τις κατατάσσει σύμφωνα με τις προτιμήσεις των ανθρώπων και συνήθως η Βελτιστοποίηση Πολιτικής Εγγύτητας (Proximal Policy Optimisation ή PPO) μεγιστοποιεί την αναμενόμενη ανταμοιβή [44]. Η Ενισχυτική Μάθηση από Ανθρώπινη Ανατροφοδότηση (Reinforcement Learning from Human Feedback ή RLHF) μειώνει τις παραισθήσεις (hallucinations) και την τοξικότητα, διατηρώντας παράλληλα τη χρησιμότητα. Αυτές οι ιδιότητες είναι κρίσιμες στις χρηματοοικονομικές ή κανονιστικές συμβουλές.

**Περιορισμένη αποκωδικοποίηση και ενίσχυση της ανάκτησης**

Τα φίλτρα ασφαλείας περιορίζουν την αναζήτηση σε policy-compliant tokens, ενώ η Παραγωγή Ενισχυμένη με Ανάκτηση (Retrieval-Augmented Generation ή RAG) εισάγει πρόσφατα ανακτημένα έγγραφα στο context-window, εξασφαλίζοντας την ακρίβεια και την επικαιρότητα των κανονιστικών αναφορών για τους χρήστες ενός εικονικού βοηθού [45].

### 2.2.3 Κύκλος ζωής LLM

Ενσωματώνοντας τα στάδια των στόχων και της ευθυγράμμισης, προκύπτει μία γραμμή παραγωγής (pipeline) όπως αυτή στην Εικόνα 2.1.




Σχήμα 2.1: *Κύκλος ζωής LLM: επιμέλεια δεδομένων, προ-εκπαίδευση, ευθυγράμμιση (SFT + RLHF), προσαρμογή τομέα και ανάπτυξη με συλλογισμό. Σχήμα 3 από Bommasani κ.α. [1].*

1. **Επεξεργασία δεδομένων** (data curation): web crawl, κώδικας, πολυγλωσσικά σώματα κειμένων, προσθήκες ειδικές για τον τομέα (π.χ οδηγοί χρήσης πλατφόρμας).
2. **Προ-εκπαίδευση** (pre-training): κατανεμημένη βελτιστοποίηση (distributed optimasation) σε χιλιάδες GPU, χρήση μικτής ακρίβειας[6] και ZeRO sharding[7] για αποδοτικότητα μνήμης.
3. **Ευθυγράμμιση** (alignment): SFT οδηγιών, RLHF και εκπαίδευση με φίλτρο ασφαλείας.
4. **Προσαρμογή τομέα** (domain adaptation): LoRA adapters[8] ή fine-tuning με αποδοτική χρήση παραμέτρων σε διαλόγους πλατφόρμας.
5. **Συμπερασματισμός** (inference): κβαντισμένα βάρη (INT8/4)[9] που παρέχονται μέσω TensorRT-LLM[10].

## 2.3 Μετρικές και σύνολα αξιολόγησης

Με τις αρχιτεκτονικές LLM και τα συστήματα εκπαίδευσης που υπάρχουν, χρειαζόμαστε βασικούς τρόπους για να μετρήσουμε αν ένα μοντέλο είναι αντικειμενικά «καλύτερο» από ένα άλλο. Αυτή η ενότητα εξετάζει τις τυπικές αυτόματες μετρικές και τις σουίτες αξιολόγησης μεγάλης κλίμακας που λειτουργούν για τέτοιου είδους κρίσεις.

### 2.3.1 Οι μετρικές BLEU και ROUGE

#### BLEU

Η μετρική Bilingual Evaluation Understudy (BLEU) συγκρίνει την έξοδο της μηχανής με μία ή περισσότερες μεταφράσεις αναφοράς μέσω τροποποιημένης ακρίβειας $n$-gram, περιορισμένης από τον αριθμό αναφορών και με ποινή σε περίπτωση συντομίας [52]. Ο γεωμετρικός μέσος όρος των ακριβειών unigram–4-gram $P_n$ δίνει:

$$\mathrm{BLEU} = BP \, \exp\Big(\sum_{n=1}^{4} w_n \log P_n\Big), \quad BP = \begin{cases} 1 & \text{αν } c > r \\ e^{1-r/c} & \text{αν } c \le r \end{cases}$$

όπου $c$ και $r$ είναι τα μήκη των υποψηφίων (candidates) και των αναφορών (references), και $w_n = \frac{1}{4}$. Το BLEU συσχετίζεται με την ποιότητα της ανθρώπινης μετάφρασης σε επίπεδο σώματος κειμένων, αλλά είναι ανεπηρέαστο από τα συνώνυμα και τη συνοχή του λόγου.

[6] Η *μικτή ακρίβεια* ή *mixed precision* αναφέρεται στην τακτική όπου τα βάρη και ενεργοποιήσεις αποθηκεύονται σε 16-bit (FP16/BF16) αντί για 32-bit, διπλασιάζοντας την απόδοση και μειώνοντας κατά το ήμισυ τη μνήμη με ελάχιστη απώλεια ακρίβειας [46].

[7] Το *Zero Redundancy Optimiser* ή *ZeRO* χωρίζει τις καταστάσεις, τις κλίσεις (gradients) και τις παραμέτρους του βελτιστοποιητή σε παράλληλες σειρές δεδομένων, δίνοντας πρόσβαση σε μοντέλα τρισεκατομμυρίων παραμέτρων σε εμπορικά προϊόντα [47].

[8] Οι *προσαρμογείς χαμηλής κατάταξης* ή *Low RAnk adapters - LoRA* εισάγουν ενημερώσεις βάρους που έχουν αποσυντεθεί κατά κατάταξη και απαιτούν μόνο $O(rd)$ παραμέτρους που μπορούν να εκπαιδευτούν, επιτρέποντας γρήγορο και φθηνό fine-tuning [48].

[9] Η κβαντοποίηση μετά την εκπαίδευση συμπιέζει τα βάρη σε ακέραιους αριθμούς 8 ή 4 bit, μειώνοντας τη μνήμη και την καθυστέρηση με <1% υποβάθμιση της ακρίβειας [49, 50].

[10] Το runtime της NVIDIA που συνενώνει πυρήνες, εφαρμόζει αραιότητα τελεστών (tensor sparcity) και μεταδίδει KV-cache για να μεγιστοποιήσει τη διακίνηση σε GPU A100/H100 [51].

#### ROUGE

Η μετρική Recall-Oriented Understudy for Gisting Evaluation (ROUGE) στοχεύει στην περίληψη. Το ROUGE-$N$ μετρά την ανάκληση (recall) των $n$-grams. Το ROUGE-L χρησιμοποιεί την μακρύτερη κοινή υποακολουθία μεταξύ υποψηφίου και αναφοράς, επιβραβεύοντας την ευχέρεια λόγου (fluency) σε επίπεδο πρότασης [53]. Σε αντίθεση με το BLEU, οι υψηλές βαθμολογίες ROUGE υποδηλώνουν ότι το μεγαλύτερο μέρος του περιεχομένου αναφοράς εμφανίζεται στην περίληψη της μηχανής, κάτι που είναι κρίσιμο για εργασίες που βασίζονται σε γεγονότα.

#### Περιορισμοί

Και οι δύο μετρικές βασίζονται στη λεξική επικάλυψη (lexical overlap) και, ως εκ τούτου, υποτιμούν τις παραφράσεις. Επιπλέον, οι αξιολογήσεις με μία μόνο αναφορά αυξάνουν τη διακύμανση. Πρόσφατες εργασίες χρησιμοποιούν μετρικές που έχουν αποκτηθεί μέσω μηχανικής μάθησης (BERTScore, COMET), οι οποίες συσχετίζονται καλύτερα με τις ανθρώπινες κρίσεις, αλλά εισάγουν μεροληψία στο μοντέλο αναφοράς [54].

### 2.3.2 Σύνολα αξιολόγησης για την κατανόηση και τη συλλογιστική

#### MMLU

Το benchmark Massive Multitask Language Understanding καλύπτει 57 θέματα προπτυχιακού επιπέδου (φυσική, νομικά, οικονομικά). Τα μοντέλα απαντούν σε ερωτήσεις πολλαπλής επιλογής. Η ακρίβεια προσεγγίζει το εύρος των εξειδικευμένων γνώσεων [55]. Οι προτροπές λίγων βολών (few-shot prompts) εξασφαλίζουν δίκαιη σύγκριση μεταξύ μοντέλων διαφορετικού μεγέθους.

#### BIG-Bench

Το Beyond the Imitation Game Benchmark (BIG-Bench) συγκεντρώνει 204 εργασίες που καλύπτουν την αριθμητική, την κοινή λογική, τη δημιουργία κώδικα και τις κοινωνικές προκαταλήψεις [56]. Η απόδοση αναφέρεται ως κανονικοποιημένη ακρίβεια σε σχέση με ανθρώπινες και τυχαίες τιμές αναφοράς. Οι αναδυόμενες δυνατότητες εμφανίζονται πάνω από 10 δισεκατομμύρια παραμέτρους, επικυρώνοντας τις τάσεις κλιμάκωσης (scaling trends).

#### Άλλες επιλογές

Το *Open LLM Leaderboard* συνδυάζει τα MMLU, HellaSwag, TruthfulQA και Winogrande. Το *AGIEval* επικεντρώνεται σε τυποποιημένες εξετάσεις (SAT, GRE). Το κεφάλαιο μεθοδολογίας (Ενότητα 3.2) βασίζεται σε αυτές τις βαθμολογίες για να επιλέξει υποψήφια βασικά μοντέλα.

Τα BLEU/ROUGE μετρούν την ευχέρεια λόγου της εξόδου των εικονικών βοηθών, ενώ το MMLU και το BIG-Bench ποσοτικοποιούν τη συλλογιστική που υποστηρίζει τις κανονιστικές συμβουλές. Η ποσοτική αξιολόγηση απαντά στο ερώτημα «Μπορεί το μοντέλο να λειτουργήσει;». Η επόμενη ενότητα εξετάζει το ερώτημα «*Πρέπει* να λειτουργήσει;», αντιμετωπίζοντας τις ηθικές και κοινωνικές επιπτώσεις.

## 2.4 Ηθικοί και κοινωνικοί περιορισμοί

Πέρα από τις τεχνικές επιδόσεις, ένας εικονικός βοηθός πρέπει να πληροί κανονιστικούς περιορισμούς σχετικά με τη δικαιοσύνη, την προστασία της ιδιωτικής ζωής, τη βιωσιμότητα και τη λογοδοσία στο πλαίσιο των κανονιστικών πλαισίων της ΕΕ.

### 2.4.1 Μεροληψία και δικαιοσύνη

Τα LLMs κληρονομούν κοινωνικές και γλωσσικές προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσης, παράγοντας στερεότυπα ή αποκλειστικά (exclusionary) αποτελέσματα [57, 58]. Η μεροληψία που σχετίζεται με το πεδίο εφαρμογής προκύπτει όταν οι ενεργειακοί κανονισμοί των κρατών μελών με υψηλό εισόδημα κυριαρχούν στο σώμα των δεδομένων, μετριάζοντας την επιρροή των τοπικών πρακτικών στην εκπαίδευση. Ο ανόμοιος αντίκτυπος μπορεί να μετρηθεί με τη «δημογραφική ισοτιμία»: δημιουργούνται συνθετικά (ψεύτικα) προφίλ χρηστών που διαφέρουν μόνο σε ένα προστατευόμενο χαρακτηριστικό (π.χ. φύλο ή ηλικία) και επαληθεύεται εάν το ποσοστό αποδοχής ή επιτυχίας του βοηθού παραμένει το ίδιο σε όλες τις ομάδες [59]. Ένα σύνολο τεχνικών, με όνομα μετριασμός (mitigation), συνδυάζει:

- *εξισορρόπηση του σώματος δεδομένων* (corpus balancing): αυξάνει την δειγματοληψία των υποεκπροσωπούμενων δημογραφικών ομάδων, έτσι ώστε οι αναλογίες των δεδομένων εκπαίδευσης να αντικατοπτρίζουν τον πληθυσμό-στόχο,
- *στόχοι εξουδετέρωσης προκαταλήψεων* (debiasing objectives): προσθέτουν όρους κανονικοποίησης που τιμωρούν τις δημογραφικές συσχετίσεις στις προβλέψεις του μοντέλου,
- *ενισχυτική μάθηση με ομαδοποιημένα δεδομένα προτιμήσεων* (reinforcement learning with grouped preference data): προσαρμόζει ένα μοντέλο ανταμοιβής στις αξιολογήσεις των ανθρώπων, ταξινομημένες ανά δημογραφική ομάδα, καθοδηγώντας την πολιτική για την εξίσωση της χρησιμότητας μεταξύ αυτών των ομάδων [44].

### 2.4.2 Ιδιωτικότητα και προστασία δεδομένων

Τα σύνολα δεδομένων σε επίπεδο κτιρίων περιέχουν προσωπικές και εμπορικές πληροφορίες που προστατεύονται από τον Γενικό Κανονισμό για την Προστασία Δεδομένων (GDPR)[11]. Τα LLMs μπορούν να απομνημονεύουν ευαίσθητες συμβολοσειρές και να τις αναπαράγουν κατά λέξη [60]. Οι τρόποι αντιμετώπισης περιλαμβάνουν (i) θόρυβο εκπαίδευσης διαφορικής ιδιωτικότητας (differential-privacy training noise) [61], (ii) ενίσχυση ανάκτησης (retrieval augmentation) που ανακτά κρυπτογραφημένα διανύσματα αντί για ακατέργαστα έγγραφα [45], και (iii) αρχεία καταγραφής ελέγχου που αποδίδουν κάθε απάντηση σε συγκεκριμένες πηγές γνώσης, ικανοποιώντας το «δικαίωμα στην εξήγηση» (right to explanation) του GDPR.

[11] *General Data Protection Regulation* ή *GDPR*: Κανονισμός (ΕΕ) 2016/679 για την προστασία των φυσικών προσώπων έναντι της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα.

### 2.4.3 Περιβαλλοντική βιωσιμότητα

Η εκπαίδευση ενός μοντέλου μπορεί να εκπέμπει από 0.65 έως και 284 τόνους $CO_2$ όταν τροφοδοτείται από το μέσο ενεργειακό δίκτυο των ΗΠΑ [62]. Οι κλιματικοί στόχοι της ΕΕ απαιτούν προγραμματισμό με γνώμονα τις εκπομπές άνθρακα, κέντρα δεδομένων που χρησιμοποιούν ανανεώσιμες πηγές ενέργειας και τεχνικές αποδοτικότητας μοντέλων (σποραδική (sparse) προσοχή, κβαντοποίηση). Αξιοσημείωτο είναι το γεγονός ότι το fine-tuning με LoRA τροποποιεί $< 1\%$ των παραμέτρων, μειώνοντας την ενέργεια κατά δύο τάξεις μεγέθους [48].

### 2.4.4 Ασφάλεια, αξιοπιστία και συμμόρφωση με τους κανονισμούς

Οι οικονομικές συμβουλές που προκύπτουν από παραισθήσεις του μοντέλου (hallucinations) μπορούν να προκαλέσουν σημαντική ζημιά. Ο νόμος της ΕΕ για την τεχνητή νοημοσύνη[12] επιβάλλει υποχρεωτική διαχείριση κινδύνων, διαφάνεια και ανθρώπινη παρέμβαση. Συγκεκριμένα, τα στοιχεία ασφάλειας είναι:

1. **Red-teaming**: προκλητικές ερωτήσεις για τον εντοπισμό τρόπων αποτυχίας [64].
2. **Αποκωδικοποίηση βάσει πολιτικής**: «μπλοκάρει» μη επιτρεπόμενο περιεχόμενο κατά τη παραγωγή απαντήσεων [65].
3. **Ανθρώπινη εποπτεία**: για όλες τις επενδυτικές συστάσεις, σύμφωνα με το άρθρο 14 του νόμου για την τεχνητή νοημοσύνη.

## 2.5 Καινοτομία και διαφοροποίηση από προηγούμενες εργασίες

Αν και η ενσωμάτωση των εικονικών βοηθών σε ψηφιακές πλατφόρμες δεν είναι κάτι καινούργιο, οι υπάρχουσες προσεγγίσεις βασίζονται συνήθως σε γενικά μοντέλα με ελάχιστη προσαρμογή στα συγκεκριμένα γλωσσικά, διαδικαστικά και κανονιστικά πλαίσια της εφαρμογής-στόχου. Προηγούμενες μελέτες, όπως αυτές των Galanakis και συν. [66] και Chaves και Gerosa [67], έχουν επισημάνει τη χρησιμότητα των VA σε βιομηχανικά και υπηρεσιακά περιβάλλοντα, αλλά συχνά παραμελούν το fine-tuning σε ροές εργασίας συγκεκριμένων τομέων ή τη βελτιστοποίηση της εμπειρίας του χρήστη μέσω βελτιώσεων που λαμβάνουν υπόψη τη διεπαφή. Η παρούσα εργασία απομακρύνεται από τις συνήθεις πρακτικές, πραγματοποιώντας εποπτευόμενο fine-tuning σε συνθετικά επεκταμένα ζεύγη προτροπών-απαντήσεων που προέρχονται απευθείας από την εσωτερική τεκμηρίωση της πλατφόρμας *ENERGATE*, δημιουργώντας έτσι ένα μοντέλο που όχι μόνο κατανοεί το τεχνικό περιεχόμενο, αλλά και ευθυγραμμίζεται στενά με τη λογική της πλατφόρμας και τους ρόλους των ενδιαφερόμενων.

Επιπλέον, η εργασία αυτή εισάγει μια ελάχιστη αλλά αποτελεσματική μορφή περιβαλλοντικής γείωσης μέσω προτροπών λέξεων-κλειδιών που προέρχονται από URL, επιτρέποντας

[12] Το παράρτημα III, σημείο 2, του νόμου της ΕΕ για την τεχνητή νοημοσύνη κατατάσσει τα «συστήματα τεχνητής νοημοσύνης που προορίζονται να χρησιμοποιηθούν ως στοιχεία ασφάλειας στη διαχείριση και τη λειτουργία [...] της παροχής νερού, φυσικού αερίου, θέρμανσης ή ηλεκτρικής ενέργειας» ως «υψηλού κινδύνου» [63].

τον διάλογο σε πραγματικό χρόνο με γνώση του πλαισίου χωρίς να τροποποιείται η αρχιτεκτονική του μοντέλου ή να προκύπτουν υψηλά κόστη συμπερασμού.

Σε αντίθεση με τους πράκτορες που βασίζονται σε κανόνες, όπως εκείνοι που κατασκευάζονται μέσω πλαισίων Agent Builder [68], οι οποίοι εξαρτώνται από αυστηρά σενάρια ροής και στατική λογική, το προτεινόμενο σύστημα επιδεικνύει ισχυρή γενίκευση από μάθηση βασισμένη σε παραδείγματα. Με αυτόν τον τρόπο, ανταποκρίνεται στις απαιτήσεις ρυθμιζόμενων, πολυγλωσσικών και δημογραφικά ετερογενών πλατφορμών όπως η *ENERGATE*, προσφέροντας μια επαναχρησιμοποιήσιμη μεθοδολογία για την ανάπτυξη χαμηλού κόστους, υψηλής πιστότητας ΕΒ σε εξειδικευμένους τομείς.

# Μέρος II

# Πρακτικό Μέρος

# Κεφάλαιο 3

# Σχεδιασμός Μεθοδολογίας

Αυτό το κεφάλαιο τυποποιεί τη μεθοδολογία για την επιλογή ενός βασικού Μεγάλου Γλωσσικού Μοντέλου (Large Language Model ή LLM) που θα γίνει fine-tune ως εικονικός πράκτορας για την πλατφόρμα ενεργειακής απόδοσης *ENERGATE*. Η Ενότητα 3.1 πραγματοποιεί μια αρχική έρευνα σε οκτώ δημόσια προσβάσιμα LLM — BLOOM, Falcon-180B, Gemini, GPT-J, GPT-NeoX, LLaMA, Vicuna-13B και XGen — επισημαίνοντας την αρχιτεκτονική τους κλίμακα, τους περιορισμούς αδειοδότησης και την καταλληλότητα του τομέα. Η Υποενότητα 3.1.1 ορίζει τους ποιοτικούς παράγοντες σύγκρισης (*Πόροι*, *Τεκμηρίωση*, *Υλοποίηση*) που καθοδηγούν την αξιολόγηση. Μια παρουσίαση μοντέλου προς μοντέλο (Υποενότητα 3.1.2) περιγράφει λεπτομερώς κάθε υποψήφιο σε σχέση με αυτούς τους παράγοντες. Η Υποενότητα 3.1.3 συγκεντρώνει τα ευρήματα σε έναν συγκριτικό πίνακα, ενώ η Υποενότητα 3.1.4 εξηγεί τη λογική της βαθμολόγησης. Η Ενότητα 3.2 συμπληρώνει την ποιοτική ανάλυση με ποσοτικά αποτελέσματα συγκριτικής αξιολόγησης που προέρχονται από το Hugging Face Open LLM Leaderboard. Η Υποενότητα 3.2.2 παρουσιάζει σε πίνακα τις βαθμολογίες σε `IFEval`, `BBH`, `GPQA`, `MuSR` και `MMLU-PRO`, ενώ η Υποενότητα 3.2.3 ερμηνεύει τις επιπτώσεις τους. Συνολικά, τα στοιχεία παρακινούν την τελική επιλογή από τα μοντέλα LLaMA προς την οικογένεια Gemini, της οποίας η cloud-native fine-tuning μέσω του Vertex AI επιλύει τους τοπικούς περιορισμούς πόρων, διατηρώντας παράλληλα κορυφαία απόδοση και τεκμηρίωση.

## 3.1 Σύγκριση LLM με βάση τα Χαρακτηριστικά

Στην παρούσα ενότητα παρουσιάζεται μια αρχική συγκριτική ανάλυση διαφόρων διακεκριμένων LLM, με στόχο τον εντοπισμό ενός κατάλληλου υποψήφιου μοντέλου για *εξειδίκευση* (fine-tuning)[1] στο πλαίσιο ενός εικονικού πράκτορα (virtual agent) για την πλατφόρμα ενεργειακής αποδοτικότητας, *ENERGATE* [8]. Τα επιλεγμένα μοντέλα (σε αλφαβητική σειρά) - BLOOM [14], Falcon-180B [69], Gemini [16], GPT-J [70], GPT-NeoX [71], LLaMA [13], Vicuna-13B [72] και XGen [73] αντιπροσωπεύουν ένα ποικίλο δείγμα ανοιχτά διαθέσιμων ή προσβάσιμων στην έρευνα LLM που έχουν αποκτήσει σημαντική έλξη στην ακαδημαϊκή κοινότητα και στην κοινότητα ανοιχτού κώδικα.

[1] Η μετάφραση του όρου *fine-tune* στα ελληνικά ποικίλλει ανάλογα με τα συμφραζόμενα, και δεν υπάρχει προς το παρόν ένα καθολικά αποδεκτό ισοδύναμο. Αποδεκτές μεταφράσεις μπορούν να είναι *εξειδίκευση (specialization)*, *περαιτέρω εκπαίδευση*, *λεπτομερή ρύθμιση* ή *προσαρμογή (adaptation)*.

Αυτά τα μοντέλα επιλέχθηκαν βάσει κριτηρίων όπως η δημόσια διαθεσιμότητα, η ποιότητα της τεκμηρίωσης, η υιοθέτηση από την κοινότητα και η συμβατότητα με την περαιτέρω προσαρμογή και ανάπτυξη. Αν και δεν είναι εξαντλητικό, το σύνολο αντικατοπτρίζει μια ισορροπία μεταξύ διαφορετικών αρχιτεκτονικών κλιμάκων, συστημάτων αδειοδότησης και στόχων σχεδιασμού (π.χ. πολυγλωσσία, αποδοτικότητα ή ευθυγράμμιση με τις ανθρώπινες οδηγίες).

Σκοπός αυτής της σύγκρισης είναι να αξιολογηθεί η πρακτική και τεχνική καταλληλότητα κάθε μοντέλου για fine-tuning και ενσωμάτωση στην προαναφερόμενη πλατφόρμα ενεργειακής αποδοτικότητας. Η ανάλυση επικεντρώνεται σε βασικά χαρακτηριστικά, όπως η λειτουργικότητα, οι υπολογιστικές απαιτήσεις, η πολυπλοκότητα της υλοποίησης, η ποιότητα της τεκμηρίωσης και η συνάφεια με την προβλεπόμενη περίπτωση χρήσης, δηλαδή η παραγωγή φυσικής γλώσσας σε ένα πολύγλωσσο, ειδικό για τον τομέα, πλαίσιο. Το αποτέλεσμα αυτής της συγκριτικής αξιολόγησης θα ενημερώσει για την επιλογή ενός βασικού μοντέλου που θα υποστεί fine-tuning για τις ειδικές ανάγκες της πλατφόρμας.

### 3.1.1 Επισκόπηση των Παραγόντων Σύγκρισης

Θα ορίσουμε αρχικά τους παράγοντες σύγκρισης, εστιάζοντας σε συγκεκριμένα ποιοτικά χαρακτηριστικά και αποφεύγοντας τεχνικές λεπτομέρειες. Έτσι, θα εστιάσουμε στις πληροφορίες που εμφανίζονται στα αντίστοιχα αποθετήρια.

#### Πόροι

Ένας από τους βασικούς παράγοντες σύγκρισης είναι το μέγεθος του μοντέλου, το οποίο συνήθως μετράται με βάση τον αριθμό των παραμέτρων. Ενώ τα μεγαλύτερα μοντέλα τείνουν να επιτυγχάνουν υψηλότερες επιδόσεις σε ένα εύρος εργασιών, συνεπάγονται επίσης σημαντικά μεγαλύτερες απαιτήσεις υπολογισμού και μνήμης τόσο κατά την εκπαίδευση όσο και κατά την εξαγωγή συμπερασμάτων.

Στο πλαίσιο της παρούσας εφαρμογής, η επιλογή ενός κατάλληλου μοντέλου απαιτεί εξισορρόπηση της απόδοσης με τους περιορισμούς των πόρων, λαμβάνοντας ιδίως υπόψη τη διαθέσιμη υποδομή και τις λειτουργικές απαιτήσεις της πλατφόρμας-στόχου. Ως εκ τούτου, τα υπερβολικά μεγάλα μοντέλα μπορεί να είναι ανεφάρμοστα, εάν μικρότερες εναλλακτικές λύσεις μπορούν να προσφέρουν συγκρίσιμα αποτελέσματα εντός αποδεκτών ορίων πόρων.

#### Τεκμηρίωση

Σημαντικός παράγοντας για την επιλογή μοντέλου είναι η διαθεσιμότητα και η ποιότητα της τεκμηρίωσης. Αυτό περιλαμβάνει επίσημα τεχνικά εγχειρίδια, κοινοτικούς πόρους (community resources) και την παρουσία δημοσιεύσεων ή ερευνητικών εργασιών με αξιολόγηση από ομοτίμους που περιγράφουν την αρχιτεκτονική του μοντέλου, τη μεθοδολογία κατάρτισης και τις προβλεπόμενες περιπτώσεις χρήσης. Η ολοκληρωμένη τεκμηρίωση διευκολύνει την κατανόηση, την ενσωμάτωση και την προσαρμογή του μοντέλου, ιδίως σε περιβάλλοντα έρευνας ή παραγωγής με περιορισμένο χρόνο ή υπολογιστικούς πόρους.

Επιπλέον, η ανάλυση αυτή εξετάζει κατά πόσον το μοντέλο είναι διαθέσιμο σε πολλαπλές παραλλαγές (π.χ. finetuned εκδόσεις ή *αποσταγμένες* (distilled) εναλλακτικές λύσεις), καθώς οι εν λόγω εκδόσεις συχνά περιλαμβάνουν βελτιώσεις στην απόδοση, την αποδοτικότητα

ή τις δυνατότητες ευθυγράμμισης. Ο βαθμός στον οποίο αυτές οι παραλλαγές τεκμηριώνονται και συντηρούνται σωστά επηρεάζει άμεσα τη χρηστικότητα και τη σημασία τους για τις *μεταεκπαιδευτικές* εργασίες (downstream tasks).

**Υλοποίηση**

Αυτός ο παράγοντας αξιολογεί τη σχετική πολυπλοκότητα της εγκατάστασης κάθε μοντέλου για δοκιμή, εκπαίδευση και ανάπτυξη. Τα μοντέλα που είναι ευκολότερα στην υλοποίηση, απαιτούν ελάχιστη διαμόρφωση, προσφέρουν υποστήριξη για τυποποιημένα πλαίσια (π.χ. PyTorch, Hugging Face Transformers [74, 75]) και περιλαμβάνουν προεκπαιδευμένα βάρη είναι πιο κατάλληλα για ταχεία πρωτοτυποποίηση και επαναληπτική ανάπτυξη.

Επιπλέον, η ευκολία υλοποίησης είναι ιδιαίτερα σημαντική σε περιβάλλοντα με περιορισμένους πόρους, όπου μοντέλα με πολύπλοκες διαδικασίες ρύθμισης ή μη τεκμηριωμένες εξαρτήσεις μπορεί να παρουσιάζουν σημαντικά εμπόδια στον πειραματισμό και την υιοθέτηση.

**Περίπτωση Χρήσης**

Η προβλεπόμενη εφαρμογή του μοντέλου παίζει καθοριστικό ρόλο στη διαδικασία επιλογής. Σε αυτό το πλαίσιο, το LLM θα χρησιμοποιηθεί αποκλειστικά για την παραγωγή κειμένου φυσικής γλώσσας σε μορφή αναγνώσιμη από τον άνθρωπο. Το μοντέλο δεν αναμένεται να εκτελεί παραγωγή κώδικα, σύνθεση εικόνων ή άλλες πολυτροπικές (multimodal) εργασίες. Κατά συνέπεια, τα μοντέλα που έχουν σχεδιαστεί ή βελτιστοποιηθεί κυρίως για αλληλεπιδράσεις με βάση το κείμενο παρουσιάζουν ιδιαίτερο ενδιαφέρον.

Επιπλέον, δεδομένου ότι τα downstream tasks περιλαμβάνουν αλληλεπιδράσεις στον τομέα της ενεργειακής αποδοτικότητας, είναι σημαντικό το βασικό μοντέλο να μπορεί να προσαρμοστεί μέσω fine-tuning για εξειδικευμένες, μη γενικής χρήσης εφαρμογές.

Εκτός αν αναφέρεται διαφορετικά, όλα τα μοντέλα που εξετάζονται σε αυτή τη σύγκριση είναι ικανά για παραγωγή κειμένου.

### 3.1.2 Επισκόπηση Μεγάλων Γλωσσικών Μοντέλων

Θα βασίσουμε την έρευνά μας στην υπηρεσία Hugging Face [76]. Το Hugging Face είναι μια κορυφαία πλατφόρμα στον τομέα της επεξεργασίας φυσικής γλώσσας που παρέχει ένα ολοκληρωμένο αποθετήριο προ-εκπαιδευμένων μοντέλων και συνόλων δεδομένων. Καλλιεργεί μια συνεργατική κοινότητα όπου ερευνητές και προγραμματιστές μπορούν να έχουν πρόσβαση σε μοντέλα τελευταίας τεχνολογίας, να συνεισφέρουν βελτιώσεις και να μοιράζονται γνώσεις. Η πλατφόρμα προσφέρει επίσης φιλικά προς το χρήστη εργαλεία και εκτενή τεκμηρίωση, διευκολύνοντας το fine-tuning των μοντέλων και την ενσωμάτωσή τους σε διάφορες εφαρμογές. Για την παρούσα έρευνα, το Hugging Face χρησιμεύει ως βασικός πόρος για την προμήθεια μοντέλων και τη συγκριτική αξιολόγηση των επιδόσεών τους.

Στην παρούσα ανάλυση, θα εξετάσουμε σε αλφαβητική σειρά τα μοντέλα: BLOOM [14], Falcon-180B [69], Gemini [16], GPT-J [70], GPT-NeoX [71], LLaMA [13], Vicuna-13B [72] και XGen [73].

### BLOOM

Το BLOOM (συντομογραφία για *BigScience Large Open-science Open-access Multilingual Language Model*) είναι ένα αυτοπαλίνδρομο γλωσσικό μοντέλο που βασίζεται σε μια στρατηγική πρόβλεψης επόμενης λέξης (next-token prediction) παρόμοια με το GPT-3 [77, 14, 78]. Ένα καθοριστικό χαρακτηριστικό του BLOOM είναι η πολύγλωσση εκπαίδευσή του: αναπτύχθηκε σε ένα τεράστιο σώμα κειμένων που καλύπτει 46 φυσικές γλώσσες και 13 γλώσσες προγραμματισμού, επιτρέποντας στο μοντέλο να προσαρμόζεται σε μια μεγάλη ποικιλία γλωσσικών πλαισίων.

#### Πόροι

Η κύρια παραλλαγή του BLOOM απαιτεί 176 δις παραμέτρους [14], αλλά έχουν διατεθεί μικρότερες εκδόσεις για να ταιριάζουν σε διαφορετικά υπολογιστικά περιβάλλοντα και περιπτώσεις χρήσης. Αυτές περιλαμβάνουν το μοντέλο με 560M παραμέτρους [79] και ενδιάμεσες εκδόσεις με 1,1B [80], 1,7B [81], 3B [82] και 7,1B [83] παραμέτρους αντίστοιχα. Αυτή η ποικιλία παρέχει στους χρήστες την ευελιξία να επιλέξουν το μέγεθος του μοντέλου που εξισορροπεί καλύτερα την απόδοση με τους περιορισμούς των πόρων.

Το επίσημο Hugging Face αποθετήριο για το BLOOM [14] χρησιμεύει ως ο κεντρικός κόμβος για την απόκτηση βαρών μοντέλων, κώδικα δειγμάτων και σημαντικών λεπτομερειών διαμόρφωσης. Προσφέρει μια κάρτα μοντέλου που εξηγεί την αρχιτεκτονική, τη διαδικασία εκπαίδευσης και τις προβλεπόμενες εφαρμογές. Παρέχονται επίσης σύνδεσμοι για κάθε παραλλαγή μεγέθους του μοντέλου, καθιστώντας εύκολη τη λήψη και τον πειραματισμό με πολλαπλές εκδόσεις.

Επιπλέον, ο οργανισμός BigScience διατηρεί ένα ιστολόγιο [78] που αφηγείται την προέλευση και την ανάπτυξη του BLOOM, μαζί με τις ιδέες που προέρχονται από την κοινότητα. Αυτό το εκτεταμένο σύνολο πόρων εξασφαλίζει μια ισχυρή βάση τόσο για αρχάριους όσο και για ειδικούς που επιθυμούν να εξερευνήσουν το BLOOM σε περιβάλλοντα έρευνας ή παραγωγής.

#### Τεκμηρίωση

Ένα από τα δυνατά σημεία του BLOOM είναι το εύρος και το βάθος της τεκμηρίωσής του, ιδιαίτερα μέσα στην υπηρεσία Hugging Face [14]. Οι χρήστες έχουν πρόσβαση σε περιεκτικές επεξηγήσεις των ρυθμίσεων του μοντέλου, της χρήσης του tokenizer[2] `BloomTokenizerFast`) και των διαφόρων κλάσεων που έχουν σχεδιαστεί για μεταγενέστερες εργασίες, όπως `BloomForCausalLM` για τη μοντελοποίηση γλώσσας, `BloomForSequenceClassification` για την ταξινόμηση κειμένου και `BloomForTokenClassification` για την αναγνώριση οντοτήτων. Υπάρχουν επίσης παραδείγματα που καταδεικνύουν τον τρόπο επίκλησης προηγμένων λειτουργιών (π.χ. απάντηση ερωτήσεων) με ελάχιστη επιβάρυνση.

Αυτή η πρακτική καθοδήγηση συμπληρώνεται από μια ερευνητική εργασία [85], η οποία εξετάζει τις θεωρητικές πτυχές της αρχιτεκτονικής του BLOOM, τους πολυγλωσσικούς στόχους και το καθεστώς εκπαίδευσης, προσφέροντας περαιτέρω σαφήνεια σε όσους στοχεύουν στην ευθυγράμμιση του μοντέλου με εξειδικευμένες ακαδημαϊκές ή βιομηχανικές

[2] Ένα *tokenizer* είναι ένα εργαλείο που μετατρέπει το ακατέργαστο κείμενο (raw text) σε μικρότερες μονάδες, όπως λέξεις, υπολέξεις ή χαρακτήρες, οι οποίες στη συνέχεια αντιστοιχίζονται σε αριθμητικές αναπαραστάσεις για επεξεργασία από το μοντέλο [84].

περιπτώσεις χρήσης.

Συλλογικά, το υλικό αυτό καλύπτει όλο το φάσμα των αναγκών των χρηστών, από τα βασικά βήματα ενσωμάτωσης έως πιο σύνθετες εργασίες όπως το fine-tuning και η αξιολόγηση του μοντέλου.

#### Υλοποίηση

Η ανάπτυξη του BLOOM βασίζεται γενικά σε τυποποιημένα συστατικά Hugging Face Transformers, καθιστώντας το μοντέλο προσβάσιμο μέσω γνωστών βιβλιοθηκών Python. Το αποθετήριο παρέχει βήμα προς βήμα οδηγίες για τη φόρτωση των προ-εκπαιδευμένων βαρών, τη μετατροπή των εισόδων κειμένου σε tokenizing, την εκτέλεση συμπερασμάτων και την εκτέλεση fine-tuning για την προσαρμογή του μοντέλου σε σύνολα δεδομένων συγκεκριμένου τομέα [14]. Αυτές οι κατευθυντήριες γραμμές περιγράφουν συστάσεις για το υλικό υπολογιστή, όπως η απαιτούμενη μνήμη GPU, και δείχνουν πώς να διαμορφώνονται οι υπερπαράμετροι (π.χ. ρυθμοί μάθησης ή μεγέθη παρτίδων) για την αντιμετώπιση συγκεκριμένων ερευνητικών ή παραγωγικών αναγκών.

Η σαφήνεια και η πληρότητα αυτών των οδηγιών διευκολύνουν την ταχεία δημιουργία πρωτοτύπων και την επαναληπτική ανάπτυξη, μειώνοντας την τεχνική επιβάρυνση που συχνά συνδέεται με την ανάπτυξη μεγάλων γλωσσικών μοντέλων. Σε συνδυασμό με τη φιλική προς το χρήστη διεπαφή του Hugging Face, η διαδικασία υλοποίησης του BLOOM είναι επαρκώς εξορθολογισμένη για μια ποικιλία περιβαλλόντων - από ακαδημαϊκά εργαστήρια με περιορισμένους πόρους έως μεγαλύτερους οργανισμούς που αναζητούν ισχυρές, πολύγλωσσες δυνατότητες παραγωγής κειμένου.

### Falcon-180B

Το Falcon-180B αποτελεί μέρος της σειράς Falcon που αναπτύχθηκε από την TII [86] και αντιπροσωπεύει ένα από τα μεγαλύτερα διαθέσιμα ανοιχτού κώδικα μοντέλα μόνο με αιτιατό αποκωδικοποιητή (casual decoder-only) [69]. Σχεδιασμένο με μια σύγχρονη αρχιτεκτονική βελτιστοποιημένη για αποδοτική εξαγωγή συμπερασμάτων, το Falcon-180B αξιοποιεί τεχνικές όπως η προσοχή πολλαπλών ερωτημάτων και αποδοτικές παραλλαγές προσοχής (efficient attention variants) όπως το `FlashAttention`. Έχει εκπαιδευτεί σε πάνω από 1 τρισεκατομμύριο tokens, με ιδιαίτερη έμφαση στο σώμα κειμένων RefinedWeb [87], γεγονός που στηρίζει τις ισχυρές επιδόσεις του σε διάφορες γλωσσικές εργασίες και κατατάσσεται σταθερά σε υψηλές θέσεις στον πίνακα κατάταξης του Open LLM Leaderboard [88].

#### Πόροι

Η οικογένεια Falcon περιλαμβάνει μικρότερες παραλλαγές (όπως Falcon-7B [89] και Falcon-40B [90]), αλλά η έκδοση 180B είναι ιδιαίτερα αξιοσημείωτη για την κλίμακα και τις δυνατότητές της, αν και απαιτεί περίπου 400 GB αποθηκευτικού χώρου [91]. Το αποθετήριο παρέχει στους χρήστες πρόσβαση σε προ-εκπαιδευμένα βάρη μοντέλων, λεπτομερείς κάρτες μοντέλων και λεπτομέρειες διαμόρφωσης (`FalconConfig, FalconModel`). Σε κάθε περίπτωση, οι μικρότερες εκδόσεις δίνουν στους ερευνητές τη δυνατότητα επιλογής ανάλογα με τους υπολογιστικούς περιορισμούς και τις απαιτήσεις της εφαρμογής τους.

#### Τεκμηρίωση

Η τεκμηρίωση που διατίθεται στο αποθετήριο του Falcon-180B είναι εκτενής και καλά δομημένη. Περιλαμβάνει εκτενείς λεπτομέρειες σχετικά με τη διαμόρφωση του μοντέλου (`FalconForCausalLM`), τη μετατροπή προσαρμοσμένων σημείων ελέγχου (`ConvertCustomCheckpoins`) και ειδικές ενότητες εργασιών για την αιτιώδη γλωσσική μοντελοποίηση, την ταξινόμηση ακολουθιών (`FalconForSequenceClassification`), την ταξινόμηση συμβόλων (`FalconForTokenClassification`) και την απάντηση ερωτήσεων (`FalconForQuestionAnswering`).

Αξιοσημείωτο είναι το γεγονός ότι (επί του παρόντος) το Falcon-180B δεν διαθέτει δημοσίευση με αξιολόγηση από ομότιμους. Η επίσημη δημοσίευσή του βρίσκεται ακόμη υπό συγγραφή [91]. Στο μεταξύ, ένα προσωρινό άρθρο [92] παρέχει πληροφορίες σχετικά με το σχεδιασμό και την απόδοση της σειράς Falcon, και μια ανάλυση του συνόλου δεδομένων RefinedWeb [87] εξηγεί τα δεδομένα εκπαίδευσης που χρησιμοποιήθηκαν. Ένα βασικό μειονέκτημα, στο πλαίσιο των πολύγλωσσων εφαρμογών, είναι ότι η εκπαίδευση του Falcon-180B καλύπτει τα αγγλικά, γερμανικά, ισπανικά, γαλλικά, ιταλικά (περιορισμένες δυνατότητες), πορτογαλικά, πολωνικά, ολλανδικά, ρουμανικά, τσέχικα και σουηδικά. Έχει περιορισμένη υποστήριξη για γλώσσες εκτός αυτής της ομάδας - κυρίως, δεν γενικεύεται καλά στα ελληνικά.

#### Υλοποίηση

Οι οδηγίες υλοποίησης για το Falcon-180B περιγράφονται σαφώς στο αποθετήριο [91]. Αυτές οι οδηγίες περιγράφουν λεπτομερώς πώς να φορτωθούν τα προ-εκπαιδευμένα βάρη, να εκτελεστεί το tokenization και να ρυθμιστεί το μοντέλο τόσο για συμπερασματολογία (inference) όσο και για fine-tuning. Επιπλέον, προσφέρουν προτάσεις σχετικά με τις απαιτήσεις του συστήματος και τις διαμορφώσεις των υπερπαραμέτρων, οι οποίες διευκολύνουν τη διαδικασία προσαρμογής του μοντέλου σε εργασίες συγκεκριμένου τομέα και στην ενσωμάτωσή του σε ευρύτερα συστήματα.

### Gemini

Το Gemini είναι μια οικογένεια πολύ ικανών πολυτροπικών μοντέλων που αναπτύχθηκε από την Google DeepMind και αντιπροσωπεύει την επόμενη γενιά μεγάλων γλωσσικών μοντέλων. Έχει σχεδιαστεί για να ενσωματώνει προηγμένη γλωσσική επεξεργασία με οπτική κατανόηση, επιτρέποντάς του να εκτελεί εργασίες που κυμαίνονται από την παραγωγή κειμένου και την απάντηση ερωτήσεων έως την ερμηνεία εικόνων [**?**, 16].

#### Πόροι

Η σειρά Gemini περιλαμβάνει διάφορες διαμορφώσεις μοντέλων που αξιοποιούν αρχιτεκτονικές τελευταίας τεχνολογίας και τεχνικές εκπαίδευσης μεγάλης κλίμακας. Μια εμπεριστατωμένη ανάλυση των γλωσσικών ικανοτήτων του Gemini αποκαλύπτει τις ανταγωνιστικές επιδόσεις του σε τυποποιημένα benchmarks, ενώ μια άλλη μελέτη αναδεικνύει τις πολυτροπικές δυνάμεις του, επιδεικνύοντας ισχυρές επιδόσεις τόσο σε εργασίες κειμένου όσο και σε εργασίες εικόνας. Επιπλέον, πηγές της βιομηχανίας αναφέρουν ότι το Gemini προσφέρεται σε πολλαπλές διαμορφώσεις: η διαμόρφωση Gemini Ultra εκτιμά-

ται σε πάνω από 1 τρισεκατομμύριο παραμέτρους, το Gemini Pro σε πάνω από 500 δισεκατομμύρια παραμέτρους και οι μικρότερες παραλλαγές - Nano-1 και Nano-2 - περιέχουν περίπου 1,8 δισεκατομμύρια και 3,25 δισεκατομμύρια παραμέτρους, αντίστοιχα [39, 93]. Τα στοιχεία αυτά είναι προσεγγιστικά και ενδέχεται να μην αντικατοπτρίζουν τις πραγματικές παραμέτρους, οι οποίες δεν έχουν δημοσιοποιηθεί.

**Τεκμηρίωση**

Η ολοκληρωμένη τεκμηρίωση για το Gemini αναδύεται μέσω πολλαπλών καναλιών. Λεπτομερείς πληροφορίες σχετικά με την αρχιτεκτονική του, το εκπαιδευτικό σχήμα και τις αξιολογήσεις των επιδόσεων παρέχονται σε πρόσφατες δημοσιεύσεις του arXiv [**?**, **?**].

**Υλοποίηση**

Αν και οι λεπτομερείς κατευθυντήριες γραμμές εφαρμογής του Gemini δεν έχουν ακόμη διαμορφωθεί, οι πρώτες ενδείξεις από τις επίσημες ανακοινώσεις της DeepMind υποδηλώνουν ότι το πλαίσιό του είναι σχεδιασμένο για απρόσκοπτη ενσωμάτωση με τα υπάρχοντα συστήματα τεχνητής νοημοσύνης. Το μοντέλο αναμένεται να υποστηρίξει την ισχυρή ανάπτυξη τόσο σε ερευνητικά όσο και σε βιομηχανικά περιβάλλοντα, με συνεχείς προσπάθειες που αποσκοπούν στην τελειοποίηση των διαδικασιών εκπαίδευσης και τελειοποίησης για τη μεγιστοποίηση της απόδοσης και της προσαρμοστικότητας.

Όπως θα αναφερθεί και παρακάτω, για την παρούσα εργασία, η ανάπτυξη ενισχύθηκε περαιτέρω με την αξιοποίηση του Vertex AI, μιας ολοκληρωμένης πλατφόρμας μηχανικής μάθησης που παρέχεται από το Google Cloud [94]. Το Vertex AI ενοποιεί ολόκληρη τη ροή εργασιών μηχανικής μάθησης - από την εισαγωγή δεδομένων και την εκπαίδευση μοντέλων έως την ανάπτυξη και την παρακολούθηση - μέσα σε ένα ενιαίο, διαχειρίσιμο περιβάλλον. Προσφέρει χαρακτηριστικά όπως AutoML, προσαρμοσμένη εκπαίδευση, επιλογές κλιμακούμενης ανάπτυξης και ολοκληρωμένα εργαλεία παρακολούθησης μοντέλων. Αυτές οι δυνατότητες απλοποιούν το fine-tuning και την ανάπτυξη μεγάλων γλωσσικών μοντέλων όπως το Gemini, επιτρέποντας την αποτελεσματική διαχείριση των πόρων, την επεκτασιμότητα και την υποστήριξη τόσο για προβλέψεις παρτίδας όσο και για προβλέψεις σε πραγματικό χρόνο.

### GPT-J

Το GPT-J είναι ένα αιτιατό γλωσσικό μοντέλο παρόμοιο με το GPT-2 [15, 95, 96], εκπαιδευμένο στο σύνολο δεδομένων Pile [97]. Αναπτύχθηκε μέσω μιας συλλογικής προσπάθειας από τους Ben Wang και Aran Komatsuzaki, και είναι ανοιχτά προσβάσιμο μέσω ενός GitHub αποθετηρίου για συνεχή ανάπτυξη και συνεισφορές της κοινότητας [98]. Παρόλο που δεν έχει δημοσίευση από ομότιμους, το GPT-J έχει ωστόσο συγκεντρώσει την προσοχή για την ισορροπία μεταξύ μεγέθους μοντέλου και επιδόσεων [70].

**Πόροι**

Η βασική έκδοση του GPT-J περιλαμβάνει 6 δις παραμέτρους, τοποθετώντας το στη μεσαία κλίμακα μεταξύ των μεγάλων γλωσσικών μοντέλων. Αυτός ο αριθμός παραμέτρων προσφέρει ένα σχετικά διαχειρίσιμο αποτύπωμα σε σύγκριση με πιο εκτεταμένα μοντέλα με εκατοντάδες δισεκατομμύρια παραμέτρων. Όπως συμβαίνει με πολλά σύγχρονα LLM, οι απαιτήσεις σε υλικό κλιμακώνονται ανάλογα με το μέγεθος του μοντέλου, απαιτώντας αρκετή

μνήμη για να διευκολυνθεί η αποτελεσματική εξαγωγή συμπερασμάτων και η fine-tuning. Παρά το μέτριο μέγεθός του, το GPT-J έχει αποδειχθεί ότι έχει ανταγωνιστικές επιδόσεις σε μια ποικιλία εργασιών - ιδίως όταν βελτιστοποιείται μέσω fine-tuning σε δεδομένα συγκεκριμένου τομέα.

**Τεκμηρίωση**

Η τεκμηρίωση GPT-J που διατίθεται στο Hugging Face repository [70] είναι εκτενής και φιλική προς το χρήστη. Παρέχει λεπτομερείς επεξηγήσεις των θεμελιωδών συστατικών, όπως `GPTJConfig` και `GPTJModel`, καθώς και των προηγμένων κλάσεων που έχουν σχεδιαστεί για downstream εργασίες, συμπεριλαμβανομένων των `GPTJForCausalLM`, `GPTJForSequenceClassification`, και `GPTJForQuestionAnswering`. Οι παράλληλες υλοποιήσεις για δημοφιλή πλαίσια βαθιάς μάθησης (TensorFlow: `TFGPTJModel`, `TFGPTJForCausalLM`, `TFGPTJForSequenceClassification`, `TFGPTJForQuestionAnswering` [99] και Flax: `FlaxGPTJModel`, `FlaxGPTJForCausalLM` [100]) αποδεικνύουν την ευελιξία του μοντέλου. Αν και δεν υπάρχει επίσημη ερευνητική εργασία που να συνοδεύει (επί του παρόντος) το GPT-J, μια ειδική ανάρτηση στο ιστολόγιο ενός από τους βασικούς προγραμματιστές παρέχει πληροφορίες σχετικά με τον σχεδιασμό και την εκπαιδευτική του προσέγγιση [101].

**Υλοποίηση**

Όπως και με άλλα μοντέλα που φιλοξενούνται στο Hugging Face, το GPT-J μπορεί εύκολα να ενσωματωθεί σε μια ποικιλία εφαρμογών. Το αποθετήριο [70] προσφέρει καθοδήγηση σχετικά με τη φόρτωση προ-εκπαιδευμένων βαρών, tokenizing κειμένου εισόδου και την εκτέλεση τόσο συμπερασμού όσο και fine-tuning. Σημειώσεις διαμόρφωσης για σημαντικές παραμέτρους (π.χ. μέγεθος παρτίδας, ρυθμός μάθησης και περιορισμοί υλικού) διευκολύνουν την προσαρμογή σε διαφορετικά περιβάλλοντα, είτε για ερευνητική διερεύνηση είτε για πιο παραγωγικές εφαρμογές. Με εύκολα προσβάσιμους κοινοτικούς πόρους και πρακτικά παραδείγματα, το GPT-J παρέχει μια σχετικά απλή πορεία για τους προγραμματιστές που αναζητούν ένα μοντέλο ενδιάμεσης κλίμακας για εργασίες παραγωγής κειμένου.

### GPT-NeoX

Το GPT-NeoX είναι ένα αυτοπαλίνδρομο γλωσσικό μοντέλο που αναπτύχθηκε από την EleutherAI [102] και εκπαιδεύτηκε στο σύνολο δεδομένων Pile [97] με σημαντική βοήθεια από την ομάδα της CoreWeave [103]. Κυκλοφορεί με ελεύθερη άδεια χρήσης και έχει ως στόχο να παρέχει μια ελεύθερα διαθέσιμη εναλλακτική λύση σε κλίμακα συγκρίσιμη με άλλα διακεκριμένα μεγάλα γλωσσικά μοντέλα [71]. Οι προγραμματιστές του υπογραμμίζουν τις ισχυρές ικανότητες συλλογισμού λίγων στιγμών (few-shot reasoning) και κερδίζει πολύ περισσότερο σε επιδόσεις όταν αξιολογείται σε πέντε βολές από ό,τι τα μοντέλα παρόμοιου μεγέθους GPT-3 [104, 105] και FairSeq [106].

**Πόροι**

Η βασική έκδοση του GPT-NeoX περιλαμβάνει 20 δις παραμέτρους [71], τοποθετώντας το σε μια ανώτερη-μέση βαθμίδα μεταξύ των σημερινών LLM. Αν και μικρότερη από ορισμένα

μοντέλα που υπερβαίνουν τα 100 δις παραμέτρους, αυτή η έκδοση εξακολουθεί να απαιτεί σημαντική μνήμη και υπολογιστικούς πόρους τόσο για το fine-tuning όσο και για την εξαγωγή συμπερασμάτων. Τα βάρη του μοντέλου και οι σχετικές διαμορφώσεις είναι εύκολα προσβάσιμα μέσω ενός ανοικτού κώδικα GitHub αποθετηρίου [107].

**Τεκμηρίωση**

Περιεκτική τεκμηρίωση για το GPT-NeoX είναι διαθέσιμη στην πλατφόρμα Hugging Face [71]. Οι χρήστες μπορούν να βρουν λεπτομερείς εξηγήσεις για διάφορες κλάσεις - όπως `GPTNeoXConfig`, `GPTNeoXTokenizerFast`, `GPTNeoXModel`, και `GPTNeoXForCausalLM`- οι οποίες υποστηρίζουν συλλογικά ένα ευρύ φάσμα εργασιών παραγωγής και ταξινόμησης κειμένου. Παρόλο που το GPT-NeoX δεν συνοδεύεται από επίσημη ερευνητική εργασία, τα ανοιχτά κοινόχρηστα αποθετήρια και οι συζητήσεις κοινότητας αντισταθμίζουν την έλλειψη ειδικής δημοσίευσης.

**Υλοποίηση**

Όπως και άλλα μοντέλα της υπηρεσίας Hugging Face, το GPT-NeoX επωφελείται από μια καθιερωμένη αλυσίδα εργαλείων για τη φόρτωση προ-εκπαιδευμένων βαρών, τη μετατροπή των εισόδων και την τροποποίηση των υπερπαραμέτρων για downstream εργασίες. Οι προγραμματιστές μπορούν να αξιοποιήσουν το επίσημο GitHub αποθετήριο [107] για να έχουν πρόσβαση στα σενάρια εκπαίδευσης και αξιολόγησης, επιτρέποντας τον πειραματισμό με fine-tuning. Αυτή η ανοικτή υποδομή προωθεί την ταχεία δημιουργία πρωτοτύπων, καθιστώντας το GPT-NeoX μια ελκυστική επιλογή για ομάδες που αναζητούν ένα γλωσσικό μοντέλο μεγάλης κλίμακας, υποστηριζόμενο από την κοινότητα.

### LLaMA

Το LLaMA (Large Language Model Meta AI) είναι μια συλλογή από θεμελιώδη γλωσσικά μοντέλα που εισήχθησαν από την Meta AI [108, 109]. Έκτοτε έχει εξελιχθεί σε πολλαπλές εκδόσεις, και συγκεκριμένα σε LLaMA2 [110, 111] και LLaMA3 [112, 113], η καθεμία με ξεχωριστά εύρη παραμέτρων και βελτιωμένες δυνατότητες. Η αρχική σειρά περιλαμβάνει μοντέλα που κυμαίνονται από 7 έως 65 δις παραμέτρους, καθιστώντας το LLaMA προσιτό τόσο στην έρευνα όσο και στη βιομηχανία, όπου οι υπολογιστικοί πόροι ποικίλλουν ευρέως.

**Πόροι**

Τα βασικά μοντέλα LLaMA καλύπτουν αριθμό παραμέτρων μεταξύ 7 δις και 65 δις [13]. Οι επόμενες εκδόσεις (LLaMA2, LLaMA3) κλιμακώνονται παρόμοια, με εκδόσεις που περιλαμβάνουν από 7 έως 70 δις παραμέτρους. Οι μεγαλύτερες παραλλαγές επιδεικνύουν επίπεδα απόδοσης συγκρίσιμα με άλλα σύγχρονα μοντέλα. Για παράδειγμα, το LLaMA-65B είναι ανταγωνιστικό με μοντέλα όπως Chinchilla-70B [114] και PaLM-540B [115], ενώ το LLaMA-13B ξεπερνά το GPT-3 (175B) [104, 105], στα περισσότερα σημεία αναφοράς.

**Τεκμηρίωση**

Υπάρχει εκτενής τεκμηρίωση για το LLaMA στο Hugging Face αποθετήριο του [13]. Εκτός από την περιγραφή γενικών συμβουλών χρήσης και πόρων, περιγράφει λεπτομερώς διάφορες κλάσεις και μεθόδους για το tokenization (`LlamaTokenizer`,

`LlamaTokenizerFast`), την αιτιώδη γλωσσική μοντελοποίηση (`LlamaForCausalLM`), την ταξινόμηση ακολουθιών (`LlamaForSequenceClassification`) και την απάντηση ερωτήσεων (`LlamaForQuestionAnswering`). Οι `LlamaConfig` και `LlamaForCausalLM` API απεικονίζουν τον τρόπο με τον οποίο οι προγραμματιστές μπορούν να κάνουν fine-tune ή να ενσωματώσουν το LLaMA σε υπάρχουσες εφαρμογές. Επιπλέον, η επίσημη ερευνητική εργασία [109] συζητά το θεωρητικό υπόβαθρο και τη μεθοδολογία εκπαίδευσης για την αρχική έκδοση του LLaMA, προσφέροντας πληροφορίες για τον αρχιτεκτονικό σχεδιασμό και τις στρατηγικές βελτιστοποίησης.

#### Υλοποίηση

Τα LLaMA GitHub αποθετήρια και οι σχετικοί πόροι δίνουν έμφαση στην πρακτική εφαρμογή και την περαιτέρω ανάπτυξη. Για παράδειγμα, ένα Google Colab αρχείο κώδικα τύπου notebook [116] δείχνει πώς να γίνει fine-tune το LLaMA χρησιμοποιώντας την τεχνική LoRA (Low-Rank Adaptation), η οποία ενεργοποιείται από τη βιβλιοθήκη HF-PEFT [117]. Ένα άλλο παρέχει οδηγίες για την εκπαίδευση και την ανάπτυξη του Open-LLaMA στο Amazon SageMaker [118], απευθυνόμενη σε χρήστες που αναζητούν κλιμακούμενες λύσεις βασισμένες στο cloud.

### Vicuna-13B

Το Vicuna-13B είναι ένας βοηθός συνομιλίας από την LMSYS [119] που αναπτύχθηκε με fine-tuning του Llama 2 σε δεδομένα συνομιλίας που συλλέχθηκαν από το ShareGPT [72, 120]. Σχεδιασμένο για να προσομοιώνει τον διαδραστικό διάλογο, το μοντέλο αυτό αξιοποιεί πραγματικές συνομιλίες χρηστών για να ενισχύσει τις συνομιλιακές του ικανότητες, καθιστώντας το έναν υποσχόμενο υποψήφιο για εφαρμογές που απαιτούν αλληλεπίδραση που μοιάζει με ανθρώπινη αλληλεπίδραση.

#### Πηγές

Η έκδοση Vicuna-13B βασίζεται σε μια αρχιτεκτονική 13 δις παραμέτρων [72]. Αυτό το μέγεθος του μοντέλου αντιπροσωπεύει έναν ισορροπημένο συμβιβασμό μεταξύ υπολογιστικής αποδοτικότητας και εκφραστικής ικανότητας, παρέχοντας ισχυρή απόδοση στη δημιουργία φυσικού, συνεκτικού διαλόγου, ενώ παραμένει προσβάσιμο για ανάπτυξη σε περιβάλλοντα με περιορισμένους πόρους.

#### Τεκμηρίωση

Για το Vicuna-13B διατίθεται ολοκληρωμένη τεκμηρίωση. Ένα λεπτομερές ερευνητικό έγγραφο [121] αξιολογεί την απόδοσή του σε μια σειρά από συγκριτικά στοιχεία συνομιλίας, και μια συνοδευτική ανάρτηση σε ιστολόγιο από την ομάδα υλοποίησης [122] προσφέρει πληροφορίες σχετικά με τη διαδικασία ανάπτυξης και τις σχεδιαστικές επιλογές. Αυτοί οι πόροι παρέχουν συλλογικά τόσο θεωρητικό υπόβαθρο όσο και πρακτική καθοδήγηση, διασφαλίζοντας ότι οι χρήστες μπορούν να κατανοήσουν και να αξιοποιήσουν αποτελεσματικά το μοντέλο.

#### Υλοποίηση

Το Vicuna-13B υποστηρίζεται από πολλαπλές πρακτικές διεπαφές. Μια διεπαφή γραμμής εντολών διευκολύνει την άμεση αλληλεπίδραση με το μοντέλο [123], ενώ τα APIs[3] (όπως αυτά που παρέχονται μέσω του OpenAI API και του Hugging Face API) επιτρέπουν την απρόσκοπτη ενσωμάτωση σε εξωτερικές εφαρμογές.

Επιπλέον, ο πηγαίος κώδικας και ο κώδικας ανάπτυξης και εγκατάστασης του μοντέλου είναι δημόσια διαθέσιμα μέσω του αποθετηρίου του στο GitHub [124], εξασφαλίζοντας τη διαφάνεια και τη συνεργασία της κοινότητας. Τέλος, μια διεπαφή επίδειξης [125] παρουσιάζει τα διαδραστικά χαρακτηριστικά του μοντέλου σε ένα φιλικό προς το χρήστη περιβάλλον.

### XGen

Το XGen είναι μια σειρά 7B μεγάλων γλωσσικών μοντέλων που αναπτύχθηκαν από την Salesforce και χρησιμοποιούν τυπικούς μηχανισμούς πυκνής προσοχής (standard dense attention mechanisms). Σύμφωνα με το ιστολόγιο της ομάδας [126], τα μοντέλα αυτά εκπαιδεύονται σε έως και 1,5 τρισεκατομμύρια tokens και βελτιώνονται σε ανοιχτά δεδομένα. Το ιστολόγιο τονίζει ότι το XGen επιτυγχάνει ανταγωνιστικές ή ανώτερες επιδόσεις σε τυποποιημένα NLP[4] σημεία αναφοράς σε σύγκριση με παρόμοια μοντέλα ανοικτού κώδικα (π.χ. MPT, Falcon, LLaMA, Redpajama, OpenLLaMA [129, 69, 13, 130, 131]).

Επιπλέον, οι αξιολογήσεις σε δείκτες αναφοράς μοντελοποίησης μακρών ακολουθιών (long sequence modeling benchmarks) καταδεικνύουν σαφή πλεονεκτήματα για τα μοντέλα που διαμορφώνονται με μήκος ακολουθίας 8K. Τέλος, το μοντέλο αρχειοθετεί εξίσου ισχυρά αποτελέσματα τόσο σε εργασίες κειμένου (`MMLU, QA`) όσο και σε εργασίες κώδικα (`HumanEval`).

#### Πόροι

Η βασική παραλλαγή, γνωστή ως XGen-7B-8K, περιλαμβάνει 7 δισεκατομμύρια παραμέτρους. Διατίθεται σε δύο διαμορφώσεις: μία εκπαιδευμένη σε ακολουθίες 4K tokens [132] και μία άλλη σε ακολουθίες 8K tokens [73]. Αυτή η ευελιξία διαμόρφωσης επιτρέπει την προσαρμογή του μοντέλου σε εργασίες που απαιτούν ποικίλα μήκη συμφραζομένων, διευρύνοντας έτσι την εφαρμογή του σε σενάρια επεξεργασίας φυσικής γλώσσας.

#### Τεκμηρίωση

Η θεμελιώδης έρευνα και οι ιδέες πίσω από το XGen παρουσιάζονται λεπτομερώς σε ένα σχετικό άρθρο [133]. Αυτή η τεκμηρίωση παρέχει μια σε βάθος συζήτηση της αρχιτεκτονικής του μοντέλου, του καθεστώτος εκπαίδευσης και των μετρήσεων απόδοσης σε εργασίες που σχετίζονται με κείμενο και κώδικα.

[3] Οι Διεπαφές Προγραμματισμού Εφαρμογών (Application Programming Interfaces - APIs) είναι τυποποιημένα σύνολα πρωτοκόλλων και εργαλείων που επιτρέπουν σε διαφορετικά συστήματα λογισμικού να επικοινωνούν μεταξύ τους. Στο πλαίσιο του Vicuna, αυτά τα APIs παρέχουν καθορισμένα τελικά σημεία (endpoints) και λειτουργίες, επιτρέποντας στους προγραμματιστές να ενσωματώσουν απρόσκοπτα τις συνομιλιακές δυνατότητες του μοντέλου σε εξωτερικές εφαρμογές με ευκολία και αποτελεσματικότητα.

[4] Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) είναι ένας τομέας της τεχνητής νοημοσύνης που επικεντρώνεται στο να επιτρέπει στους υπολογιστές να κατανοούν, να ερμηνεύουν και να παράγουν την ανθρώπινη γλώσσα [127, 128].

#### Υλοποίηση

Οδηγίες για την ανάπτυξη και την ανάπτυξη του μοντέλου XGen-7B-8K παρέχονται στο σχετικό αποθετήριο [73]. Αυτές οι οδηγίες περιλαμβάνουν διαδικασίες για fine-tuning, τη διαχείριση των διαμορφώσεων μήκους ακολουθίας και την ενσωμάτωση του μοντέλου σε μεγαλύτερες σωληνώσεις NLP. Οι διαθέσιμοι πόροι διευκολύνουν τον πρακτικό πειραματισμό, διευκολύνοντας τους ερευνητές και τους προγραμματιστές να αξιοποιήσουν τις δυνατότητες του XGen τόσο σε ακαδημαϊκά όσο και σε παραγωγικά περιβάλλοντα.

### 3.1.3 Συγκριτικός Πίνακας Χαρακτηριστικών

Στην παρούσα ενότητα παρουσιάζεται ένας συγκριτικός πίνακας για τη συστηματική αξιολόγηση και σύγκριση των επιλεγμένων LLM με βάση τους παράγοντες σύγκρισης χαρακτηριστικών που αναλύθηκαν προηγουμένως. Ο πρωταρχικός στόχος αυτού του πίνακα είναι να παρέχει μια σαφή και συνοπτική επισκόπηση των δυνατών σημείων και των περιορισμών κάθε μοντέλου, διευκολύνοντας έτσι μια τεκμηριωμένη απόφαση για την επακόλουθη τελειοποίηση και ανάπτυξη. Ο πίνακας εξετάζει τρεις βασικές διαστάσεις: διαθεσιμότητα πόρων (όπως υποδεικνύεται από το εύρος των παραμέτρων του μοντέλου), ποιότητα τεκμηρίωσης και ευκολία υλοποίησης. Συνοψίζοντας αυτά τα χαρακτηριστικά σε δομημένη μορφή, η ανάλυση αυτή επιδιώκει να γεφυρώσει το χάσμα μεταξύ των θεωρητικών μετρήσεων επίδοσης και των πρακτικών ζητημάτων υλοποίησης.

#### Ορισμός Κριτηρίων Αξιολόγησης

Τα κριτήρια αξιολόγησης ταυτίζονται με τους παράγοντες σύγκρισης χαρακτηριστικών που αναλύθηκαν στην Υποενότητα *3.1.1 Επισκόπηση των Παραγόντων Σύγκρισης* και αναλύθηκαν ξεχωριστά για κάθε μοντέλο στην Υποενότητα *3.1.2 Επισκόπηση Μεγάλων Γλωσσικών Μουτέλων*. Η στήλη *Πόροι* απαριθμεί τα διαθέσιμα μεγέθη μοντέλων, παρουσιάζεται περιγραφικά και δεν υπόκειται σε βαθμολόγηση. Από την άλλη, για να εξασφαλιστεί μια αυστηρή και αναπαραγώγιμη αξιολόγηση των επιλεγμένων LLM, τα άλλα δύο κριτήρια αξιολόγησης - *Ποιότητα Τεκμηρίωσης* και *Ευκολία Υλοποίησης* - βαθμολογούνται ποσοτικά σε μια κλίμακα Likert [134] από 1 έως 5. Συγκεκριμένα:

#### Ποιότητα Τεκμηρίωσης

Αυτό το κριτήριο αξιολογεί την έκταση και τη σαφήνεια της επίσημης τεκμηρίωσης και των συμπληρωματικών πόρων του μοντέλου. Για την βαθμολόγηση του κριτηρίου αυτού θα εξεταστούν:

- *Ερευνητικές εργασίες*: Η παρουσία επίσημων ερευνητικών εργασιών ή ισοδύναμων επιστημονικών δημοσιεύσεων που περιγράφουν την αρχιτεκτονική του μοντέλου και τη μεθοδολογία εκπαίδευσης.

- *Επίσημη βιβλιογραφία*: Η ποιότητα και η πληρότητα των καρτών μοντέλου, των κατευθυντήριων γραμμών και των τεχνικών εγχειριδίων που παρέχονται σε επίσημα αποθετήρια.

- *Συμπληρωματικοί πόροι*: Διαθεσιμότητα πρόσθετων πόρων, όπως κοινοτικά σεμινάρια, παραδείγματα κώδικα και φόρουμ χρηστών που διευκολύνουν την κατανόηση και τη χρήση του μοντέλου.
- *Περιεκτικότητα και σαφήνεια*: Ο βαθμός στον οποίο η τεκμηρίωση εξηγεί με σαφήνεια τις διαδικασίες διαμόρφωσης, την αντιμετώπιση προβλημάτων και τα βήματα ενσωμάτωσης.

Σύμφωνα με την κλίμακα Likert, κάθε μοντέλο θα βαθμολογείται από 1 έως 5 με:

- *5*: Εκτενής, καλά οργανωμένη τεκμηρίωση που υποστηρίζεται από πολλαπλές δημοσιεύσεις με αξιολόγηση από ομοτίμους και άφθονα πρακτικά παραδείγματα.
- *4*: Καλή τεκμηρίωση με σαφείς οδηγίες και αρκετά παραδείγματα, αν και ορισμένες προχωρημένες πτυχές ενδέχεται να απαιτούν περαιτέρω διευκρινίσεις.
- *3*: Βασική τεκμηρίωση που καλύπτει βασικές λειτουργίες- οι χρήστες μπορεί να χρειαστεί να συμπληρώσουν με κοινοτικούς πόρους.
- *2*: Περιορισμένη τεκμηρίωση που παρέχει ελάχιστη καθοδήγηση, που πιθανόν να οδηγήσει σε προκλήσεις στην ενσωμάτωση του μοντέλου.
- *1*: Αραιή ή ανεπαρκής τεκμηρίωση με σημαντικά κενά σε τεχνικές λεπτομέρειες.

**Ευκολία Υλοποίησης**

Αυτό το κριτήριο αξιολογεί την πρακτικότητα της ανάπτυξης κάθε μοντέλου, δίνοντας έμφαση στη διαθεσιμότητα οδηγιών ενσωμάτωσης και στην πολυπλοκότητα που συνεπάγεται η προσαρμογή του μοντέλου για συγκεκριμένες εργασίες. Η αξιολόγηση βασίζεται στα εξής:

- *Οδηγίες υλοποίησης*: Η διαθεσιμότητα οδηγών βήμα προς βήμα για την ανάπτυξη του μοντέλου εντός τυποποιημένων πλαισίων (π.χ. Hugging Face Transformers, PyTorch).
- *Κοινοτική υποστήριξη*: Η παρουσία πηγών που συνεισφέρει η κοινότητα, συμπεριλαμβανομένων εκπαιδευτικών οδηγιών, σημειωματάριων παραδειγμάτων και φόρουμ αντιμετώπισης προβλημάτων.
- *Δυσκολία διαμόρφωσης*: Η εγγενής πολυπλοκότητα της διαμόρφωσης του μοντέλου, των fine-tuning διαδικασιών και της διαχείρισης εξαρτήσεων.
- *Κατευθυντήριες γραμμές για υλικό πόρους*: Σαφήνεια σχετικά με τις απαιτήσεις υλικού και τις συστάσεις για βέλτιστη ανάπτυξη.

Σύμφωνα με την κλίμακα Likert, κάθε μοντέλο θα βαθμολογείται από 1 έως 5 με:

- *5*: Η ανάπτυξη είναι ιδιαίτερα απλοποιημένη με ολοκληρωμένους οδηγούς και ισχυρή υποστήριξη από την κοινότητα, ελαχιστοποιώντας το κόστος διαμόρφωσης.
- *4*: Η ανάπτυξη είναι γενικά απλή, αν και ενδέχεται να υπάρχουν μικρές προκλήσεις που απαιτούν πρόσθετες διευκρινίσεις.

- *3*: Το μοντέλο μπορεί να αναπτυχθεί με μέτρια προσπάθεια- ορισμένες διαμορφώσεις μπορεί να απαιτούν σημαντικές προσαρμογές ή πρόσθετη έρευνα.

- *2*: Η ανάπτυξη είναι πολύπλοκη και ανεπαρκώς τεκμηριωμένη, γεγονός που μπορεί να εμποδίσει την πρακτική εφαρμογή.

- *1*: Το μοντέλο είναι πολύ δύσκολο να αναπτυχθεί λόγω έλλειψης σαφών οδηγιών και εκτεταμένων προκλήσεων διαμόρφωσης.

#### Συγκριτικός Πίνακας

Με βάση τα κριτήρια αξιολόγησης που περιγράφονται παραπάνω, ο Πίνακας 3.1 παρουσιάζει τον συγκριτικό πίνακα των επιλεγμένων LLM.

| **LLM** | **Πόροι** | **Τεκμηρίωση** | **Υλοποίηση** |
|---|---|---|---|
| LLaMA | 7B → 65B | 5 | 5 |
| BLOOM | 560M → 176B | 5 | 4 |
| Falcon | 7B / 40B / 180B | 3 | 2 |
| Gemini | 1.8B → 1T | 5 | 5 |
| XGen | 7B | 3 | 3 |
| GPT-NeoX | 20B | 3 | 4 |
| GPT-J | 6B | 3 | 4 |
| Vicuna | 13B | 4 | 4 |

Πίνακας 3.1: *Συγκριτική επισκόπηση των LLM: Διαθεσιμότητα Πόρων, Ποιότητα Τεκμηρίωσης και Ευκολία Ανάπτυξης*

### 3.1.4 Σχολιασμός Βαθμολογίας

Ακολουθεί λεπτομερής επεξήγηση των βαθμολογιών κάθε μοντέλου, με βάση τα κριτήρια αξιολόγησης που ορίζονται στην Υποενότητα 3.1.3 Ορισμός Κριτηρίων Αξιολόγησης.

- **LLaMA**: Το LLaMA λαμβάνει βαθμολογία 5/5 τόσο στην ποιότητα τεκμηρίωσης όσο και στην ευκολία υλοποίησης. Τα επίσημα αποθετήριά του συμπληρώνονται από εκτεταμένη έρευνα με αξιολόγηση από ομοτίμους και από πληθώρα κοινοτικών πόρων. Οι λεπτομερείς οδηγίες ενσωμάτωσης και τελειοποίησης το καθιστούν ιδιαίτερα προσιτό.

- **BLOOM**: Το BLOOM βαθμολογείται με 5/5 για την ποιότητα της τεκμηρίωσης και με 4/5 για την ευκολία ανάπτυξης. Παρόλο που η τεκμηρίωσή του είναι πλήρης και καλύπτει πολλαπλά μεγέθη μοντέλων, η ποικιλία των παραλλαγών μπορεί να εισάγει πολυπλοκότητα κατά τη διαδικασία ανάπτυξης.

- **Falcon**: Το Falcon λαμβάνει βαθμολογία 3/5 για την ποιότητα της τεκμηρίωσης και 2/5 για την ευκολία ανάπτυξης. Ενώ είναι επαρκώς τεκμηριωμένο για βασική χρήση, περιορισμοί όπως η μειωμένη πολυγλωσσική υποστήριξη και η υψηλότερη πολυπλοκότητα ενσωμάτωσης οδηγούν σε χαμηλότερη ευκολία ανάπτυξης.

- **Gemini**: Το Gemini λαμβάνει βαθμολογία 5/5 και στις δύο κατηγορίες. Παρά την εκτεταμένη τεκμηρίωση που παρέχεται μέσω της Vertex AI - η οποία μπορεί να φαίνεται κάπως διασκορπισμένη λόγω του όγκου των διαθέσιμων εργαλείων - η συνολική κάλυψη είναι ολοκληρωμένη. Η πλήρης πλατφόρμα της Vertex AI απλοποιεί σημαντικά την ενσωμάτωση και την ανάπτυξη, δικαιολογώντας τις κορυφαίες βαθμολογίες.

- **XGen**: Το XGen βαθμολογείται με 3/5 τόσο στην ποιότητα τεκμηρίωσης όσο και στην ευκολία ανάπτυξης. Η τεκμηρίωσή του είναι επαρκής, αλλά όχι τόσο εμπεριστατωμένη όσο αυτή των κορυφαίων μοντέλων, και οι οδηγίες ανάπτυξης, αν και λειτουργικές, μπορεί να αποτελέσουν πρόκληση λόγω της ανάγκης διαχείρισης διαφορετικών διαμορφώσεων μήκους ακολουθίας.

- **GPT-NeoX**: Το GPT-NeoX λαμβάνει βαθμολογία 3/5 για την ποιότητα της τεκμηρίωσης και 4/5 για την ευκολία εγκατάστασης. Η τεκμηρίωσή του καλύπτει τις βασικές πτυχές, αλλά συχνά απαιτεί συμπληρωματικούς κοινοτικούς πόρους, ενώ η διαδικασία εγκατάστασης είναι γενικά απλή.

- **GPT-J**: Το GPT-J βαθμολογείται με 3/5 για την ποιότητα τεκμηρίωσης και 4/5 για την ευκολία ανάπτυξης. Παρόλο που η τεκμηρίωσή του καλύπτει επαρκώς τις βασικές λειτουργίες, οι χρήστες ενδέχεται να χρειαστεί να ανατρέξουν σε πόρους της κοινότητας για πιο προηγμένες ρυθμίσεις. Η διαδικασία ανάπτυξης υποστηρίζεται σχετικά καλά.

- **Vicuna**: Το Vicuna λαμβάνει βαθμολογία 4/5 και στις δύο κατηγορίες. Η τεκμηρίωσή του συνδυάζει αποτελεσματικά τους επίσημους πόρους με τις γνώσεις της κοινότητας και παρέχει διάφορες διεπαφές ανάπτυξης, όπως εργαλεία γραμμής εντολών και API.

## 3.2 Σύγκριση LLM με βάση την Απόδοση

Εκτός από τα κριτήρια αξιολόγησης που επιλέχθηκαν, οι δείκτες αναφοράς προσφέρουν ένα αντικειμενικό μέσο για την αξιολόγηση γλωσσικών μοντέλων σε τυποποιημένες εργασίες. Το Hugging Face Open LLM Leaderboard παρέχει ένα ενοποιημένο πλαίσιο για την αξιολόγηση των LLM σε διάφορα κριτήρια αναφοράς, συμπεριλαμβανομένων της απάντησης ερωτημάτων, της παραγωγής κειμένου και της γλωσσικής μοντελοποίησης.

### 3.2.1 Επισκόπηση του Open LLM Leaderboard

Το Hugging Face Open LLM Leaderboard συγκεντρώνει δείκτες αναφοράς για γλωσσικά μοντέλα ανοικτού κώδικα χρησιμοποιώντας το Eleuther AI Language Model Evaluation Harness, ένα ενοποιημένο πλαίσιο για τη δοκιμή γεννητικών γλωσσικών μοντέλων σε μεγάλο αριθμό διαφορετικών εργασιών αξιολόγησης [88, 135].

Περιλαμβάνει έξι βασικά σημεία αναφοράς που έχουν σχεδιαστεί για να ελέγχουν διάφορες πτυχές: `IFEval`, `BBH`, `GPQA`, `MuSR` και `MMLU-PRO`. Κάθε δείκτης αναφοράς αξιολογεί διαφορετικές διαστάσεις της απόδοσης του μοντέλου, όπως η προσήλωση στις οδηγίες, η ικανότητα συλλογισμού και η γενική γνώση. Η ανάλυση αυτών των βαθμολογιών σε

ρυθμίσεις μηδενικών και ελάχιστων βολών (zero and few shot) παρέχει ποσοτικοποιημένα στοιχεία που συμπληρώνουν την προηγούμενη ποιοτική σύγκριση, προσφέροντας μια ολοκληρωμένη εικόνα της αποτελεσματικότητας κάθε μοντέλου σε ένα ευρύ φάσμα εργασιών. Συγκεκριμένα, οι δείκτες αναφοράς είναι:

- **IFEval (Instruction-Following Evaluation)**: Το IFEval έχει σχεδιαστεί για να αξιολογεί μια βασική ικανότητα των LLM: την ικανότητα να ακολουθούν οδηγίες φυσικής γλώσσας, χρησιμοποιώντας ένα σύνολο επαληθεύσιμων οδηγιών. Παρακάμπτει το υψηλό κόστος και την υποκειμενικότητα των ανθρώπινων αξιολογήσεων παρέχοντας ένα αναπαραγώγιμο μέτρο σύγκρισης. Το σύνολο δεδομένων αποτελείται από περίπου 500 προτροπές που καλύπτουν 25 διαφορετικούς τύπους επαληθεύσιμων οδηγιών. Αυτή η απλή προσέγγιση διευκολύνει τη συνεπή αυτόματη αξιολόγηση των επιδόσεων παρακολούθησης οδηγιών [136].

- **BBH (Big Bench Hard)**: Το BBH είναι ένα επιμελημένο υποσύνολο 23 ιδιαίτερα απαιτητικών εργασιών από την ευρύτερη σουίτα Big-Bench. Αυτές οι εργασίες επιλέχθηκαν ειδικά επειδή προηγούμενες αξιολογήσεις έδειξαν ότι τα γλωσσικά μοντέλα γενικά αποδίδουν σε αυτές κάτω από το μέσο ανθρώπινο επίπεδο. Ο συγκεκριμένος δείκτης αναφοράς υπογραμμίζει τη σημασία της συλλογιστικής πολλαπλών βημάτων, όπου τεχνικές όπως ερωτήματα αλυσιδωτών σκέψεων έχουν δείξει ότι βελτιώνουν την απόδοση. Συνεπώς, χρησιμεύει ως αυστηρό τεστ για την αξιολόγηση των αναδυόμενων ικανοτήτων συλλογιστικής των LLM [137].

- **MATH (Mathematics Aptitude Test of Heuristics)**: Το σύνολο δεδομένων MATH περιλαμβάνει 12.500 απαιτητικά μαθηματικά προβλήματα διαγωνιστικού επιπέδου που συνοδεύονται από πλήρεις λύσεις βήμα προς βήμα. Σκοπός του είναι να αξιολογήσει τις ικανότητες μαθηματικής συλλογιστικής και επίλυσης προβλημάτων των LLM, μια ικανότητα όπου τα τρέχοντα μοντέλα εξακολουθούν να δυσκολεύονται. Παρά την αύξηση του μεγέθους και της πολυπλοκότητας του μοντέλου, τα αποτελέσματα δείχνουν ότι η ισχυρή μαθηματική συλλογιστική παραμένει ασύλληπτη. Το σημείο αναφοράς υπογραμμίζει την ανάγκη για αλγοριθμικές καινοτομίες πέρα από την απλή κλιμάκωση [138].

- **GPQA (Graduate-Level Google-Proof Q&A)**: Το GPQA είναι ένα απαιτητικό σύνολο δεδομένων πολλαπλών επιλογών που περιέχει 448 ερωτήσεις που αναπτύχθηκαν από ειδικούς του τομέα της βιολογίας, της φυσικής και της χημείας. Οι ερωτήσεις έχουν σχεδιαστεί για να είναι εξαιρετικά δύσκολες αλλά «Google-proof»[5], με τις επιδόσεις σε επίπεδο εμπειρογνωμόνων να φτάνουν συνήθως μόνο περίπου στο 65% της ακρίβειας. Ακόμη και προηγμένα μοντέλα όπως το GPT-4 έχουν επιτύχει μόνο μέτρια αποτελέσματα σε αυτό τον δείκτη αναφοράς. Ο «Google-proof» χαρακτήρας του καθιστά το GPQA ένα πολύτιμο εργαλείο για τη δοκιμή των ορίων των σημερινών συστημάτων τεχνητής νοημοσύνης σε εξειδικευμένους γνωστικούς τομείς [139].

[5]Ερωτήσεις που είναι επαληθεύσιμες με απεριόριστη πρόσβαση στο διαδίκτυο.

- **MuSR (Multistep Soft Reasoning)**: Το MuSR αξιολογεί LLM σε πολλαπλών βημάτων εργασίες ήπιας συλλογιστικής που παρουσιάζονται ως αφηγήσεις φυσικής γλώσσας. Προκαλεί τα μοντέλα με μακρά, πολύπλοκα προβλήματα, όπως μυστήρια δολοφονίας ή περίπλοκες εργασίες βελτιστοποίησης, που απαιτούν την ενσωμάτωση της συλλογιστικής με ανάλυση συμφραζομένων σε μεγάλες αποστάσεις (long distance context analysis). Το σύνολο δεδομένων δημιουργείται με τη χρήση ενός νευροσυμβολικού αλγορίθμου συνθετικής μετατροπής σε φυσική μορφή, εξασφαλίζοντας τόσο την πολυπλοκότητα όσο και τον ρεαλισμό. Αυτό το μέτρο σύγκρισης αποκαλύπτει σημαντικά κενά στις τρέχουσες τεχνικές συλλογιστικής, ακόμη και όταν χρησιμοποιείται η προτροπή της αλυσίδας σκέψης [140].

- **MMLU-Pro (Massive Multitask Language Understanding Professional)**: Το MMLU-Pro είναι μια βελτιωμένη έκδοση του Massive Multitask Language Understanding benchmark, επανασχεδιασμένη ώστε να ενσωματώνει πιο δύσκολες ερωτήσεις έντονου συλλογισμού. Επεκτείνει το σύνολο επιλογών απάντησης από τέσσερις σε δέκα και φιλτράρει τις τετριμμένες ή θορυβώδεις ερωτήσεις, αυξάνοντας έτσι τη συνολική δυσκολία. Το σύνολο δεδομένων παρουσιάζει σημαντική πτώση της ακρίβειας σε σύγκριση με το αρχικό MMLU, αναδεικνύοντας την αυξημένη διακριτική του ικανότητα. Επιπλέον, παρουσιάζει μεγαλύτερη σταθερότητα σε διαφορετικά είδη προτροπών, υποδεικνύοντας ένα πιο στιβαρό πλαίσιο αξιολόγησης [141].

### 3.2.2 Συγκριτικός Πίνακας Αποδόσεων

Στην παρούσα ενότητα παρουσιάζεται ένας συγκριτικός πίνακας για τη συστηματική αξιολόγηση και σύγκριση των επιλεγμένων LLM με βάση την βαθμολογία στους δείκτες απόδοσης που αναλύθηκαν προηγουμένως. Είναι σημαντικό να σημειωθεί ότι τα μοντέλα XGen, GPT-J και Gemini δεν περιλαμβάνονται σε αυτή την αξιολόγηση του συγκριτικού πίνακα είτε επειδή δεν έχουν μετρηθεί ακόμη είτε, στην περίπτωση του Gemini, λόγω περιορισμών που σχετίζονται με τη μη ανοιχτού κώδικα φύση του. Τα αποτελέσματα συνοψίζονται στον Πίνακα 3.2.

| **LLM** | **IFEval** | **BBH** | **GPQA** | **MuSR** | **MMLU-PRO** |
|---|---|---|---|---|---|
| LLaMA | 84.92 | 55.41 | 16.55 | 17.09 | 47.21 |
| BLOOM | 13.22 | 4.04 | 1.9 | 1.92 | 1.16 |
| Falcon | 24.54 | 17.22 | 0 | 5.16 | 14.02 |
| Gemini | - | - | - | - | - |
| XGen | - | - | - | - | - |
| GPT-NeoX | 25.87 | 4.93 | 0 | 2.82 | 1.73 |
| GPT-J | - | - | - | - | - |
| Vicuna-13B | 33.44 | 7.49 | 2.35 | 4.09 | 13.81 |

Πίνακας 3.2: *Συγκριτική επισκόπηση των LLM με βάση την απόδοση τους σε διάφορους δείκτες αναφοράς (IFEval, BBH, GPQA, MuSR, MMLU-PRO)*

### 3.2.3 Σχολιασμός Βαθμολογίας

Ο συγκριτικός πίνακας παρέχει πολύτιμες πληροφορίες σχετικά με τα σχετικά πλεονεκτήματα και τους περιορισμούς των γλωσσικών μοντέλων που αξιολογήθηκαν. Θα πρέπει να σημειωθεί ότι οι βαθμολογίες που αναφέρονται αντιστοιχούν σε συγκεκριμένες παραλλαγές μοντέλων, συγκεκριμένα:

- LLaMA 3.1-70b για το LLaMA,
- 7b1 για το BLOOM,
- 40b για Falcon, και
- Έκδοση 1.3 για το Vicuna-13B.

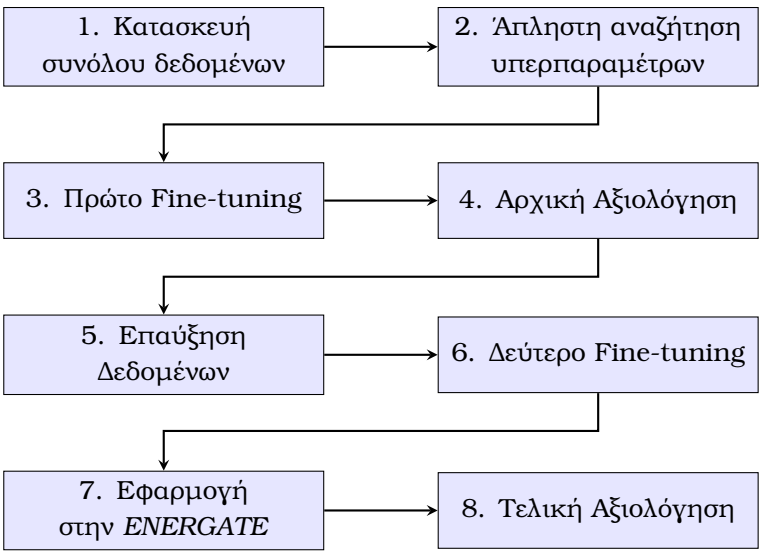
Συνολικά, το LLaMA επιδεικνύει ανώτερες επιδόσεις σε πολλαπλά σημεία αναφοράς σε σύγκριση με το BLOOM και άλλα μοντέλα, γεγονός που ενισχύεται περαιτέρω από την ολοκληρωμένη τεκμηρίωση και την ευκολία ενσωμάτωσής του. Με βάση αυτά τα ευρήματα, μια αρχική στρατηγική ήταν η υιοθέτηση του LLaMA ως βασικού προτύπου για περαιτέρω fine-tuning και ενσωμάτωση. Ωστόσο, μεταγενέστερες πρακτικές εκτιμήσεις - κυρίως οι υπολογιστικές προκλήσεις που σχετίζονται με το fine-tuning του LLaMA σε τοπικό υλικό- οδήγησαν σε επαναξιολόγηση της προσέγγισης.

Υπό το πρίσμα αυτών των περιορισμών, επιλέχτηκε σαν εναλλακτική το Gemini, το οποίο, παρά το γεγονός ότι δεν συμπεριλήφθηκε στην αξιολόγηση των δεικτών αναφοράς, προσφέρει ισχυρή υποστήριξη για το fine-tuning σε πλατφόρμες που βασίζονται στο νέφος, όπως η Vertex AI. Αυτή η προσέγγιση όχι μόνο μετριάζει την ανάγκη για σημαντικούς τοπικούς πόρους GPU, αλλά αξιοποιεί επίσης την εκτεταμένη σουίτα εργαλείων και τεκμηρίωσης ενσωμάτωσης της Vertex AI, εξασφαλίζοντας έτσι μια ομαλότερη διαδικασία ανάπτυξης.

Κεφάλαιο 4

# Εφαρμογή Μεθοδολογίας

Το κεφάλαιο αυτό παρουσιάζει την πρακτική εφαρμογή της μεθοδολογίας που ορίστηκε στο Κεφάλαιο 3 Σχεδιασμός Μεθοδολογίας. Ξεκινά με μια περιγραφή της αρχικής προσπάθειας εκτέλεσης μιας τοπικής διαδικασίας βελτιστοποίησης χρησιμοποιώντας LLaMA και Ollama, η οποία τελικά εγκαταλείφθηκε λόγω τεχνικών περιορισμών. Στη συνέχεια, το κεφάλαιο τεκμηριώνει την πλήρη διαδικασία ανάπτυξης και βελτιστοποίησης του επιλεγμένου μοντέλου Gemini χρησιμοποιώντας το Google Cloud Vertex AI. Η φάση υλοποίησης φαίνεται στο Σχήμα 4.1. Ξεκινώντας από τη συλλογή και προετοιμασία των δεδομένων εκπαίδευσης, η ροή εργασίας προχώρησε στη ρύθμιση του περιβάλλοντος στο Vertex AI, ακολουθούμενη από fine-tuning. Η διαδικασία περιελάμβανε στρατηγικές άπληστης εύρεσης κατάλληλων παραμέτρων, επαύξησης δεδομένων για την ενίσχυση του μοντέλου ώστε να αναγνωρίζει την τοποθεσία του χρήστη, καθώς και ολοκληρωμένα βήματα αξιολόγησης και η εξαγωγή συμπερασμάτων με βάση δειγματοληψία. Κάθε φάση ενημέρωνε τις προσαρμογές της επόμενης, δημιουργώντας έναν βρόχο ανατροφοδότησης μεταξύ των αποτελεσμάτων του μοντέλου και των βελτιώσεων των δεδομένων εκπαίδευσης.




Σχήμα 4.1: *Ροή εργασίας υλοποίησης για την εκπαίδευση και την αξιολόγηση μοντέλων*

## 4.1 Πειραματική εφαρμογή LLaMA

Η αρχική στρατηγική υλοποίησης επικεντρώθηκε στην ανάπτυξη μιας τοπικά εκτελέσιμης έκδοσης του μοντέλου LLaMA. Δεδομένης της ανοιχτής πρόσβασης σε διάφορες παραλλαγές του LLaMA και της δυνατότητας βελτιστοποίησής τους σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων, αυτή η προσέγγιση κρίθηκε βιώσιμη σε αρχικό στάδιο.

### 4.1.1 Τοπική ανάπτυξη μέσω Ollama

Η πρώτη προσπάθεια πραγματοποιήθηκε χρησιμοποιώντας το Ollama, ένα ελαφρύ περιβάλλον σχεδιασμένο για τοπική εκτέλεση LLM [142]. Η διαδικασία εγκατάστασης περιελάμβανε τη λήψη του επίσημου προγράμματος εγκατάστασης Ollama, την επαλήθευση της εγκατάστασης μέσω εντολών τερματικού και, στη συνέχεια, τη λήψη της επιθυμητής παραλλαγής LLaMA [143]. Μετά τη λήψη των παραγόμενων αρχείων του μοντέλου, δοκιμάστηκε η βασική αλληλεπίδραση με βάση διαλόγου.

Για την εκτέλεση σεναρίων (scripts) Python, δημιουργήθηκε ένα εικονικό περιβάλλον με όνομα `chatbot`. Εγκαταστάθηκαν τα απαιτούμενα πακέτα (`langchain`, `langchain-ollama` και `ollama`) και δημιουργήθηκε ένα δοκιμαστικό σενάριο `main.py` (Παράρτημα Β΄ Δοκιμή τοπικής χρήσης LLaMA). Παρά την επιτυχή τοπική συμπερασματολογία, η ενσωμάτωση του συστήματος αποκάλυψε ασυνέπειες στη διαδρομή.

### 4.1.2 Περιορισμοί υλικού και έρευνα κβαντοποίησης

Το σύστημα που χρησιμοποιήθηκε για την τοπική ανάπτυξη ήταν εξοπλισμένο με επεξεργαστή Intel Core i5-12400 12ης γενιάς, GPU NVIDIA RTX 3050 (8 GB VRAM) και 8 GB RAM. Παρά τις μέτριες δυνατότητες, οι δοκιμές έδειξαν ότι ακόμη και μοντέλα 7 δισεκατομμυρίων παραμέτρων δεν μπορούσαν να λειτουργήσουν με συνέπεια χωρίς σοβαρή κβαντοποίηση[1] (Πίνακας 4.1 Απαιτήσεις VRAM μοντέλων LlaMA για αποκλειστική εκτέλεση σε GPU με κβαντοποίηση 4-Bit). Αυτό οδήγησε σε μια αξιολόγηση μεθόδων κβαντοποίησης όπως GGML (για συμπεράσματα βασισμένα σε CPU) και GPTQ (για εκτέλεση με επιτάχυνση GPU), οι οποίες μειώνουν την ακρίβεια για να επιτρέψουν τη λειτουργία με χαμηλή μνήμη [144]:

- **GGML**: Το GGML (GPT-Generated Model Language) είναι μια μέθοδος κβαντοποίησης που έχει σχεδιαστεί για τη βελτιστοποίηση των LLMs για συμπεράσματα CPU, ιδιαίτερα σε συστήματα χωρίς ισχυρές GPU. Εστιάζει στη μείωση του μεγέθους του μοντέλου διατηρώντας παράλληλα την απόδοση, καθιστώντας το κατάλληλο για συσκευές με περιορισμένη VRAM.

- **GPTQ**: Το GPTQ (Generalized Post-Training Quantization) είναι μια άλλη προσέγγιση κβαντοποίησης που στοχεύει στην αποτελεσματική εξαγωγή συμπερασμάτων σε GPU. Μειώνει το μέγεθος του μοντέλου και το υπολογιστικό φορτίο, επιτρέποντας σε

[1] Η *κβαντοποίηση* (quantization) αναφέρεται στη διαδικασία μείωσης της ακρίβειας των βαρών του μοντέλου (π.χ. από 32-bit κινητής υποδιαστολής σε 8-bit ή 4-bit αναπαραστάσεις) προκειμένου να μειωθεί η χρήση μνήμης και οι υπολογιστικές απαιτήσεις, επιτρέποντας σε μεγάλα μοντέλα να λειτουργούν αποτελεσματικά σε περιορισμένο υλικό [144].

μεγαλύτερα μοντέλα να εκτελούνται σε GPU με λιγότερη VRAM. Το GPTQ είναι ιδιαίτερα επωφελές για χρήστες με GPU NVIDIA, καθώς αξιοποιεί την επιτάχυνση GPU για να βελτιώσει την απόδοση.

Πολλά μοντέλα, συμπεριλαμβανομένων των Vicuna, WizardLM, Airoboros και Guanaco, προσδιορίστηκαν ως πιθανοί υποψήφιοι λόγω των 4-bit κβαντοποιημένων παραλλαγών τους και της συμβατότητάς τους με διαμορφώσεις 7B. Παρόλα αυτά, το διαθέσιμο υλικό δεν ήταν διαθέσιμο για να υποστηρίξει τα εν λόγω μοντέλα.

| **LLM Model** | **Model Size** | **Min. VRAM Req.** | **GPU Examples** |
|---|---|---|---|
| Llama 2 & Llama 3.1 | 8B | 6GB | RTX 3060, RTX 4060, GTX 1660, 2060, AMD 5700 XT, RTX 3050 |
| Llama 2 | 13B | 10GB | AMD 6900 XT, RTX 2060 12GB, 3060 12GB, RTX 4070, RTX 2080, 2080 Ti |
| LLama | 33B | 20GB | RTX 3080 20GB, RTX 4000 Ada, A4500, A5000, 3090, 4090, 6000, Tesla V100, Tesla A40 |
| Llama 2 & Llama 3.1 | 70B | 40GB | A100 40GB, 2×3090, 2×4090, A40, RTX 6000, 8000 |
| Llama 3.1 | 405B | 232GB | 10×3090, 10×4090, 6×A100 40GB, 3×H100 80GB |

Πίνακας 4.1: *Απαιτήσεις VRAM μοντέλων LlaMA για αποκλειστική εκτέλεση σε GPU με κβαντοποίηση 4-Bit*

### 4.1.3 Υποδομή και οικονομικοί περιορισμοί

Η περαιτέρω διερεύνηση εναλλακτικών λύσεων βασισμένων σε API, συμπεριλαμβανομένων των υπηρεσιών LLaMAIndex και LLaMA API, δεν είχε επιτυχία. Ενώ τα endpoints προσέφεραν online συμπερασμούς και fine-tuning, επέβαλαν περιορισμούς χρήσης βάσει πιστώσεων που απαιτούσαν χρηματικό υπόλοιπο. Αυτός ο περιορισμός ήταν ασυμβίβαστος με τους όρους του ερευνητικού έργου, το οποίο απαιτούσε δωρεάν και ανοιχτά εργαλεία χωρίς τέλη χρήσης.

### 4.1.4 Συμπεράσματα της τοπικής ανάπτυξης LLaMA

Συνοψίζοντας, αν και το οικοσύστημα LLaMA προσέφερε διάφορα μοντέλα ανοιχτού κώδικα και επιτεύχθηκε μερική επιτυχία στη διαμόρφωση ενός τοπικού χρόνου εκτέλεσης, οι απαιτήσεις υλικού για αξιόπιστη συμπερασματολογία και ειδικά για fine-tuning υπερέβαιναν τις δυνατότητες του διαθέσιμου συστήματος. Η ανάγκη για χρηματικό πιστωτικό όριο σε δημόσια API απέκλεισε περαιτέρω τις λύσεις εξωτερικής φιλοξενίας. Αυτοί οι παράγοντες κατέστησαν το LLaMA ακατάλληλο για τους στόχους υλοποίησης της παρούσας διπλωματικής εργασίας και κατέστησαν αναγκαία τη μετάβαση σε μια εναλλακτική λύση cloud-native, η οποία περιγράφεται στην επόμενη ενότητα.

## 4.2 Εφαρμογή στο Google Cloud Vertex AI

Μετά την αποτυχημένη προσπάθεια τοπικής ανάπτυξης και προσαρμογής ενός βοηθού βασισμένου στο LLaMA, η εργασία στράφηκε στην πλατφόρμα Vertex AI του Google Cloud ως εναλλακτική λύση. Αυτή η μετάβαση καθοδηγήθηκε τόσο από τεχνικούς περιορισμούς όσο και από τη συγκριτική αξιολόγηση των γλωσσικών μοντέλων και των περιβαλλόντων που πραγματοποιήθηκε στο Κεφάλαιο 3 Σχεδιασμός Μεθοδολογίας. Τα μοντέλα Gemini, διαθέσιμα μέσω του Vertex AI, αναδείχθηκαν ως μια πρακτική επιλογή λόγω της απόδοσής τους, της εκτενούς τεκμηρίωσης και της ενσωματωμένης υποστήριξης για βελτιστοποίηση μέσω διαδικτυακών εργαλείων.

Το Vertex AI προσφέρει πρόσβαση σε μια σειρά γενετικών υπηρεσιών τεχνητής νοημοσύνης, συμπεριλαμβανομένης της οικογένειας μοντέλων Gemini, με πρόσβαση βάσει πιστώσεων μέσω δωρεάν προφίλ χρήστη. Αυτό έκανε την πλατφόρμα κατάλληλη για πειραματισμό εντός των περιορισμών ενός ακαδημαϊκού ερευνητικού έργου. Στόχος ήταν η χρήση του Vertex AI για το fine-tuning ενός LLM ώστε να λειτουργεί ως εικονικός βοηθός ενσωματωμένος στην πλατφόρμα *ENERGATE*.

### 4.2.1 Επιλογή πλατφόρμας και αρχική ρύθμιση

Αρχικά, δημιουργήθηκαν λογαριασμοί Google Cloud, που μπορούσαν να χρησιμοποιηθούν για την εξερεύνηση και τη λειτουργία των υπηρεσιών του Vertex AI. Τα μοντέλα Gemini που είναι διαθέσιμα στην πλατφόρμα υποστηρίζουν διαφορετικούς τρόπους λειτουργίας και ροές εργασίας βελτιστοποίησης. Μέσα από μια εκτενή ανασκόπηση των διαθέσιμων εργαλείων και της τεκμηρίωσης, εντοπίσαμε δύο διαφορετικές προσεγγίσεις σχετικές με τον στόχο της παρούσας εργασίας:

- **Agent Builder**: ένα εργαλείο βασισμένο γραφική διεπαφή για τη δημιουργία πρακτόρων (agents) βασισμένων σε κανόνες χρησιμοποιώντας εγχειρίδια (playbooks) και υπό όρους λογική [68].
- **Supervised Fine-Tuning**: μια μέθοδος για fine-tuning μοντέλων Gemini σε σύνολα δεδομένων που παρέχονται από τον χρήστη χρησιμοποιώντας επισημασμένα παραδείγματα εκπαίδευσης [145].

Αφού εξετάστηκαν οι αντίστοιχες τεχνικές απαιτήσεις, οι δυνατότητες και η ευθυγράμμιση με την περίπτωσή της εργασίας, επιλέχθηκε η ροή εργασίας *Supervised Fine-tuning*. Το *Agent Builder* απορρίφθηκε λόγω της άκαμπτης δομής του προτύπου, της περιορισμένης προσαρμοστικότητάς του στη λογική του συγκεκριμένου τομέα και της εξάρτησής του από προκαθορισμένες απαντήσεις αντί για μάθηση βασισμένη σε μοντέλα. Στις επόμενες ενότητες περιγράφεται λεπτομερώς αυτή η αξιολόγηση, ακολουθούμενη από το σχεδιασμό και την υλοποίηση της ροής εποπτευόμενου finetuning.

### 4.2.2 Σύγκριση *Agent Builder* και *Fine-Tuning*

Μόλις ολοκληρώθηκε η αρχική ρύθμιση στο Vertex AI, διερευνήθηκαν οι διαθέσιμες επιλογές για την υλοποίηση ενός εικονικού βοηθού μέσω της σουίτας γενετικής τεχνητής

νοημοσύνης της πλατφόρμας. Οι δύο κύριες διαδρομές που προσφέρει το Google Cloud για τη δημιουργία προσαρμοσμένων γλωσσικών πρακτόρων ήταν το *Agent Builder* και το *Supervised Fine-tuning*. Κάθε μέθοδος ακολουθούσε μια θεμελιωδώς διαφορετική προσέγγιση στον τρόπο με τον οποίο ορίζεται και εκτελείται η συμπεριφορά του βοηθού.

Το *Agent Builder* είναι μια λύση χωρίς κώδικα που έχει σχεδιαστεί για τη δημιουργία πρακτόρων συνομιλίας βάσει κανόνων. Βασίζεται στον χειροκίνητο ορισμό των *playbooks*, δηλαδή δομημένων συνόλων προκαθορισμένων προθέσεων, ενεργειών και μεταβάσεων. Κάθε playbook αντιπροσωπεύει μια λογική ροή του διαλόγου του πράκτορα και οι απαντήσεις επιλέγονται με βάση την υπό όρους λογική. Σε αυτήν την περίπτωση, το μοντέλο δεν βελτιστοποιείται ούτε ενημερώνεται, αλλά τα αποτελέσματά του περιορίζονται από ρητές οδηγίες και στατικά δέντρα αποφάσεων [68].

Το *Supervised Fine-tuning*, από την άλλη πλευρά, περιλαμβάνει την εκπαίδευση ενός γενετικού μοντέλου σε ένα επιλεγμένο σύνολο δεδομένων με παραδείγματα εισόδου-εξόδου. Το μοντέλο εσωτερικεύει τη συμπεριφορά μέσω ενημερώσεων κλίσης, με αποτέλεσμα ένα νέο σύνολο παραμέτρων που καθοδηγούν τις μελλοντικές απαντήσεις του [145]. Αυτή η μέθοδος επιτρέπει την προσαρμογή σε συγκεκριμένους τομείς και υποστηρίζει πιο ευέλικτες και ευαίσθητες στο περιβάλλον αλληλεπιδράσεις.

Μετά την ανάλυση των δυνατοτήτων και των περιορισμών και των δύο προσεγγίσεων, αποφασίστηκε ότι το *Agent Builder* δεν ήταν κατάλληλο για τους σκοπούς της εργασίας. Η στατική αρχιτεκτονική του στερείται της ικανότητας γενίκευσης που απαιτείται για τη διαχείριση ερωτημάτων φυσικής γλώσσας σχετικά με την πλατφόρμα *ENERGATE*. Επιπλέον, η μετατροπή των τεσσάρων οδηγών χρήστη σε μεμονωμένα playbooks με λογική διακλάδωσης θα ήταν υπερβολικά περίπλοκη και αντιπαραγωγική, δεδομένου του στόχου της εκπαίδευσης ενός διαλογικού πράκτορα που μαθαίνει από την επίδειξη και όχι από τον προγραμματισμό.

Ως καταλληλότερη μέθοδος επιλέχθηκε το εποπτευόμενο fine-tuning. Αυτή επιτρέπει τον άμεσο έλεγχο της συμπεριφοράς του μοντέλου μέσω της μάθησης βάση παραδειγμάτων, υποστηρίζει τη φυσική γλωσσική ποικιλομορφία τόσο στις ερωτήσεις όσο και στις απαντήσεις και προσφέρει μεγαλύτερη επεκτασιμότητα και προσαρμοστικότητα στο εξελισσόμενο περιεχόμενο της πλατφόρμας.

#### Απόρριψη του *Agent Builder*

Για να κατανοηθεί καλύτερα η αρχιτεκτονική του *Agent Builder*, εξετάστηκε ένα από τα προκαθορισμένα πρότυπα του Google Cloud: έναν εικονικό βοηθό για το Τμήμα Οχημάτων (Department of Motor Vehicles ή DMV). Αυτό το πρότυπο αποτελείται από ένα σύνολο προκαθορισμένων playbooks, καθένα από τα οποία αντιπροσωπεύει ένα συγκεκριμένο σενάριο αλληλεπίδρασης. Τα τέσσερα βασικά playbooks σε αυτό το παράδειγμα είναι: *DMV Steering*, *Renew Driver's License*, *Book Appointment* και *Fallback*.

Κάθε playbook περιέχει αυστηρές οδηγίες και μεταβάσεις με κώδικα. Για παράδειγμα, στο playbook *DMV Steering*, ο βοηθός έχει ρητή οδηγία να μην απαντά απευθείας στην ερώτηση του χρήστη, αλλά να τον ανακατευθύνει σε ένα από τα δύο playbooks προορισμού, ανάλογα με το περιεχόμενο του αιτήματος.

Αυτή η άκαμπτη, μη προσαρμοστική δομή απεικονίζεται στο ακόλουθο απόσπασμα από

το πρότυπο:

```
Playbook: DMV Steering

Goal:
Your goal is to collect the customer's request and route them to the
correct service.

Instructions:
- Greet the customer.
- DO NOT help the user directly.
- ALWAYS transfer to one of the following:
  - $PLAYBOOK: Renew Driver's License
  - $PLAYBOOK: Book Appointment
  - $PLAYBOOK: Fallback
```

Τα επόμενα playbooks ακολουθούν παρόμοια διαδικαστική λογική, απαιτώντας ακριβή πεδία εισαγωγής δεδομένων από τον χρήστη (π.χ. αριθμός άδειας, ημερομηνία λήξης, ημερομηνία γέννησης) και καλώντας εργαλεία ψευδοκώδικα (π.χ. `$TOOL: dmv_renew_tool`) για να καθορίσουν την επόμενη ενέργεια. Αξίζει να σημειωθεί ότι αυτά τα εργαλεία δεν αποτελούν μέρος της επεξεργασίας του γλωσσικού μοντέλου, αλλά αντιπροσωπεύουν σταθερές κλήσεις API ή ανακατευθύνσεις.

Αυτό το πρότυπο που βασίζεται σε κανόνες έρχεται σε έντονη αντίθεση με τον στόχο αυτής της διπλωματικής εργασίας: να δημιουργηθεί ένας εικονικός βοηθός ικανός για ευφραδή, διάλογο με επίγνωση των συμφραζομένων βασισμένο στην τεκμηρίωση της πλατφόρμας *ENERGATE*. Η πολυπλοκότητα και η αλληλεπικάλυψη των ρόλων των χρηστών στην πλατφόρμα (π.χ. αλληλεπικάλυψη λειτουργιών μεταξύ Ιδιοκτητών Κτιρίων (Building Owners) και Υπευθύνων Υλοποίησης (Implementors)) καθιστούν αδύνατη την εκ των προτέρων προγραμματισμένη κωδικοποίηση κάθε διαδρομής λήψης αποφάσεων.

Αν και είναι τεχνικά δυνατό να μετατραπεί κάθε οδηγός χρήστη σε ένα αντίστοιχο σύνολο playbooks, αυτό θα ήταν αντίθετο με τον σχεδιασμό των βοηθών LLM που βασίζονται στη μάθηση και το fine-tuning. Κατά συνέπεια, το πλαίσιο *Agent Builder* απορρίφθηκε ως ασυμβίβαστο με τους στόχους της παρούσας εργασίας.

### 4.2.3 Εποπτευόμενο fine-tuning του Gemini

Αφού η προσέγγιση *Agent Builder* απορρίφθηκε, εξετάστηκε η επιλογή του *Supervised Fine-tuning* των μοντέλων Gemini μέσω του Vertex AI. Αυτή η επιλογή συνάδει με τον στόχο της δημιουργίας ενός ευέλικτου, ευαίσθητου στο περιβάλλον βοηθού, εκπαιδεύοντάς τον απευθείας σε παραδείγματα συγκεκριμένων εργασιών. Σύμφωνα με την επίσημη τεκμηρίωση, το εποπτευόμενο fine-tuning επιτρέπει στο μοντέλο να μάθει τις επιθυμητές συμπεριφορές από τα επισημασμένα δεδομένα, κάτι που είναι ιδιαίτερα κατάλληλο για ταξινόμηση, εξαγωγή οντοτήτων, περίληψη και απάντηση σε ερωτήσεις συγκεκριμένου τομέα [145].

Τα μοντέλα Gemini που υποστηρίζουν εποπτευόμενο fine-tuning κατά τη στιγμή της σύνταξης του παρόντος είναι τα `gemini-2.0-flash-001` και `gemini-2.0-flash-lite-001`. Και

τα δύο είναι βελτιστοποιημένα για χαμηλή καθυστέρηση και οικονομικά αποδοτική εξαγωγή συμπερασμάτων. Η περίπτωση χρήσης της παρούσας εργασίας – η απάντηση σε δομημένες ερωτήσεις που προέρχονται από την τεκμηρίωση της *ENERGATE* – εμπίπτει πλήρως στο εύρος των εφαρμογών για τις οποίες προορίζονται.

#### Αξιολόγηση εγγράφων εισόδου

Ένας από τους διαθέσιμους τύπους εισόδου δεδομένων για εποπτευόμενο fine-tuning είναι η εκπαίδευση βάσει εγγράφων (document tuning). Αυτό επιτρέπει στους χρήστες να ανεβάζουν αρχεία PDF (Portable Document File) ως πηγή γνώσης. Ωστόσο, αυτή η μέθοδος εισάγει σημαντικούς περιορισμούς. Πρώτον, τα PDF μετατρέπονται σε tokens με μορφή εικόνας, πράγμα που σημαίνει ότι το κείμενο δεν αναλύεται σημασιολογικά, αλλά ως είσοδος τύπου εικόνας, γεγονός που υποβαθμίζει την κατανόηση και την ακρίβεια του μοντέλου. Δεύτερον, η διαχείριση εγγράφων του Gemini κληρονομεί την τιμολόγηση και τους περιορισμούς της εισόδου με βάση εικόνες, καθιστώντας το μια υποβέλτιστη επιλογή για κείμενο μεγάλης κλίμακας.

Συγκεκριμένα, το πλαίσιο βελτιστοποίησης επιβάλλει τους ακόλουθους περιορισμούς στα δεδομένα PDF:

- Μέγιστο 300 σελίδες ανά παράδειγμα
- Μέγιστο 4 αρχεία PDF ανά παράδειγμα
- Μέγιστο μέγεθος αρχείου: 20MB

Επιπλέον, η τεκμηρίωση της πλατφόρμας προειδοποιεί για τη χρήση εικονοποιημένων PDF, τονίζει τη χρήση κειμένου που μπορεί να αναγνωστεί από μηχανή και επισημαίνει την κακή απόδοση του μοντέλου στον χωρικό συλλογισμό ή στην ερμηνεία της διάταξης των εγγράφων [146]. Δεδομένης της δομής των οδηγών *ENERGATE*, οι οποίοι είναι εξ ολοκλήρου κειμενικής φύσεως, και του στόχου της ακρίβειας της σημασιολογικής απόκρισης, επιλέξαμε να μετατρέψουμε το περιεχόμενο σε μορφή κειμένου.

#### Fine-tuning βάσει κειμένου

Η εισαγωγή κειμένου είναι ο προτιμώμενος τύπος δεδομένων για fine-tuning όταν η βάση γνώσεων αποτελείται από κείμενα ειδικού τομέα, οδηγίες ή συχνές ερωτήσεις. Το Vertex AI απαιτεί δεδομένα εκπαίδευσης και επικύρωσης σε μορφή JSON Lines (JSONL [147]), όπου κάθε γραμμή κωδικοποιεί μια πλήρη σειρά διαλόγου ή ένα ζεύγος οδηγίας-απάντησης.

Κάθε δείγμα εκπαίδευσης αποτελείται από δύο κύρια στοιχεία: το `systemInstruction`, το οποίο μπορεί προαιρετικά να ορίζει την προσωπικότητα ή τον ρόλο του βοηθού, και μια λίστα `contents` που περιέχει τον διάλογο προτροπής-απόκρισης μεταξύ του χρήστη και του μοντέλου. Ακολουθεί ένα απλοποιημένο παράδειγμα:

```
{
  "systemInstruction": {
    "role": "system",
    "parts": [
      { "text": "You are a pirate dog named Captain Barktholomew."
        }
```

```
    ]
  },
  "contents": [
    { "role": "user", "parts": [ { "text": "Hi" } ] },
    { "role": "model", "parts": [ { "text": "Argh! What brings ye
       to my ship?" } ] }
  ]
}
```

Στην περίπτωση της παρούσας εργασίας, το `systemInstruction` παραλείφθηκε. Κάθε δείγμα αποτελείτο από μια ερώτηση χρήστη και την αντίστοιχη απάντηση που εξήχθη ή παραφράστηκε από τους οδηγούς χρήσης και τις συχνές ερωτήσεις (Frequently Asked Questions ή FAQ) της *ENERGATE*. Αυτό επέτρεψε να προσομοιωθούν ρεαλιστικές αλληλεπιδράσεις της πλατφόρμας, διατηρώντας παράλληλα αυστηρό έλεγχο της ακρίβειας των απαντήσεων και της ευθυγράμμισης του τομέα.

### 4.2.4 Μετατροπή της βάσης γνώσεων σε JSONL

Για την μετατροπή του μοντέλου Gemini, το κειμενικό περιεχόμενο της πλατφόρμας *ENERGATE* έπρεπε να μετατραπεί σε δομημένα δεδομένα εκπαίδευσης σε μορφή JSONL. Το αρχικό υλικό περιελάμβανε τέσσερις επίσημους οδηγούς χρήστη — έναν για κάθε ρόλο ενδιαφερόμενου μέρους (Building Owners, Implementors, Financiers και Public Bodies[2]) — καθώς και την ενότητα «Συχνές Ερωτήσεις» (FAQ) της *ENERGATE*. Τα έγγραφα αυτά περιγράφουν τις λειτουργίες, τις διαδικασίες και τις ροές εργασίας της πλατφόρμας που σχετίζονται με κάθε τύπο χρήστη.

Η διαδικασία μετατροπής περιελάμβανε τη χειροκίνητη εξαγωγή δηλωτικών τμημάτων κειμένου από τους οδηγούς και τη συσχέτιση καθενός με μια σχετική ερώτηση σε φυσική γλώσσα. Για παράδειγμα, μια παράγραφος που εισήγαγε την πλατφόρμα *ENERGATE* συνδυάστηκε με πολλαπλές διατυπώσεις της ερώτησης «*What is the ENERGATE platform?*»[3]. Αυτή η μορφή ερώτησης-απάντησης αντικατοπτρίζει ρεαλιστικές αλληλεπιδράσεις των χρηστών, περιορίζοντας παράλληλα την απόδοση του βοηθού σε ακριβείς, τεκμηριωμένες απαντήσεις.

Τα παραδείγματα που προέκυψαν κωδικοποιήθηκαν στο απαιτούμενο σχήμα JSONL, χρησιμοποιώντας τη δομή που περιγράφεται στην Ενότητα 4.2.3 Fine-tuning βάσει κειμένου. Κάθε γραμμή στο σύνολο δεδομένων αντιπροσώπευε μια σειρά διαλόγου, με τον χρήστη να παρέχει μια προτροπή και το μοντέλο να απαντά με το κατάλληλο κείμενο που εξήχθη. Το επαναλαμβανόμενο ή δομικά περιττό περιεχόμενο σε όλους τους οδηγούς απομακρύνθηκε προσεκτικά για να αποφευχθεί η υπερβολική εκπροσώπηση και να διατηρηθεί η ισορροπία μεταξύ των ρόλων των χρηστών.

[2] *Ιδιοκτήτες Κτιρίων*, *Υπεύθυνοι Υλοποίησης*, *Χρηματοδότες* και *Δημόσιοι Φορείς*, αντίστοιχα
[3] Μτφ: Τι είναι η πλατφόρμα ENERGATE;

**Δημιουργία και επαύξηση συνόλου δεδομένων εκπαίδευσης και επικύρωσης**

Το σύνολο δεδομένων χωρίστηκε σε υποσύνολα εκπαίδευσης (training) και επικύρωσης (validation) σύμφωνα με τις οδηγίες του Vertex AI. Αρχικά, δημιουργήθηκαν 120 μοναδικά ζεύγη χρήστη-μοντέλου από το περιεχόμενο της *ENERGATE*. Δεδομένου του σχετικά μικρού μεγέθους του σώματος κειμένων, εφαρμόστηκε αύξηση των παραδειγμάτων για να αυξηθεί ο όγκος και η γλωσσική ποικιλία. Για κάθε βασική προτροπή, δημιουργήθηκαν πέντε εναλλακτικές εκφράσεις, με χρήση του ChatGPT [148]. Αυτές ήταν σημασιολογικά ισοδύναμες αλλά συντακτικά διαφορετικές, αντανακλώντας τους διαφορετικούς τρόπους με τους οποίους ένας χρήστης μπορεί να θέσει την ίδια ερώτηση. Κάθε παραλλαγή διατήρησε την ίδια έξοδο του μοντέλου.

Μια επιπλέον έκτη παραλλαγή της προτροπής τοποθετήθηκε στο σύνολο επικύρωσης, με αποτέλεσμα αναλογία 5:1 μεταξύ των παραδειγμάτων εκπαίδευσης και επικύρωσης. Αυτή η μέθοδος δημιούργησε ένα σύνολο εκπαίδευσης 600 παραδειγμάτων και ένα σύνολο επικύρωσης 120 παραδειγμάτων, συνολικά 720 εγγραφές JSONL. Παρακάτω παρουσιάζεται ένα ενδεικτικό παράδειγμα:

```
{
  "contents": [
    { "role": "user", "parts": [ { "text": "What is the ENERGATE
       platform?" } ] },
    { "role": "model", "parts": [ { "text": "The ENERGATE Energy
       Efficiency (EE) Marketplace (hereinafter referred to as
       ENERGATE platform) is an ICT platform that aims to
       facilitate communication and financial deal closure between
       market actors involved in energy efficiency projects for
       buildings. ..." } ] }
  ]
}
```

Με άλλα λόγια, για το συγκεκριμένο παράδειγμα, το σύνολο εκπαίδευσης θα έχει τις εξής παραλλαγές ερωτήσεων:

1. `What is the ENERGATE platform?`
2. `Can you tell me about the ENERGATE platform?`
3. `Could you explain what the ENERGATE platform is?`
4. `What does the ENERGATE platform do?`
5. `I'd like to know more about the ENERGATE platform. What is it?`

Ενώ το σύνολο επικύρωσης θα είχε μία έκτη παραλλαγή:

1. `Please provide an overview of the ENERGATE platform.`

Φυσικά, η απάντηση και στις έξι παραλλαγές έμεινε ίδια (`The ENERGATE Energy Efficiency (EE) Marketplace...`). Το μέγεθος του συνόλου δεδομένων επιλέχθηκε ώστε να παραμείνει εντός της κλίμακας των παραδειγμάτων συνόλων δεδομένων που παρέχονται στην επίσημη τεκμηρίωση του Vertex AI, όπου είναι συνηθισμένα σύνολα εκπαίδευσης 500–1000 παραδειγμάτων [149]. Η στρατηγική ελεγχόμενης αναπαραγωγής χρησίμευσε

επίσης για την ενθάρρυνση της γενίκευσης χωρίς να θυσιάζεται η ακρίβεια, διασφαλίζοντας ότι ο βοηθός θα λειτουργεί αξιόπιστα σε όλες τις συνήθεις ερωτήσεις των χρηστών.

## 4.3 Εκτέλεση fine-tuning μοντέλου

Το τελικό στάδιο της υλοποίησης περιελάμβανε την εκτέλεση της διαδικασίας εποπτευόμενου fine-tuning χρησιμοποιώντας την υποδομή fine-tuning του Vertex AI. Αυτό απαιτούσε τη μεταφόρτωση των προετοιμασμένων συνόλων δεδομένων στον αποθηκευτικό χώρο του cloud, τη διαμόρφωση των παραμέτρων της εργασίας βελτιστοποίησης και την εκκίνηση της διαδικασίας είτε μέσω της γραφικής διεπαφής είτε προγραμματικά μέσω του Python SDK. Σε αυτήν την ενότητα περιγράφεται κάθε βήμα της διαδικασίας, συμπεριλαμβανομένης της παρακολούθησης των εργασιών και των μετρήσεων αξιολόγησης που χρησιμοποιούνται για την αξιολόγηση της απόδοσης του μοντέλου σε πολλαπλές εκτελέσεις βελτιστοποίησης.

#### Φόρτωση συνόλου δεδομένων στο GCS

Το Vertex AI απαιτεί τα δεδομένα εκπαίδευσης και επικύρωσης να αποθηκεύονται στο Google Cloud Storage (GCS) πριν από την έναρξη οποιασδήποτε εργασίας ρύθμισης. Το GCS οργανώνει τα αρχεία σε δοχεία (buckets), τα οποία λειτουργούν ως φακέλους για την αποθήκευση αρχείων.

Για να ανεβάσουμε τα σύνολα δεδομένων JSONL, δημιουργήσαμε πρώτα ένα νέο bucket μέσω του Google Cloud Console. Αυτό γίνεται μεταβαίνοντας στο Cloud Storage Browser στο [150] και επιλέγοντας την επιλογή «Create Bucket». Κάθε bucket πρέπει να έχει ένα μοναδικό όνομα και να είναι συνδεδεμένο με μια συγκεκριμένη περιοχή (π.χ. `us-central1`). Για λόγους συνέπειας, τόσο τα αρχεία εκπαίδευσης όσο και τα αρχεία επικύρωσης τοποθετήθηκαν στον ίδιο bucket, αλλά διαχωρίστηκαν με βάση το όνομα του αρχείου. Αυτές οι διαδρομές αρχείων αναφέρθηκαν αργότερα κατά τη διαμόρφωση της εργασίας:

- `gs://bucket_epu_thesis_2025/TrainDataset.jsonl`
- `gs://bucket_epu_thesis_2025/ValidationDataset.jsonl`

Η διεπαφή επιτρέπει μεταφορτώσεις με τη χρήση διεπαφής ή μέσω CLI με το βοηθητικό πρόγραμμα γραμμής εντολών `gsutil`.

### 4.3.1 Εκτέλεση Fine-Tuning Εργασιών

Μόλις τα σύνολα δεδομένων μεταφορτώθηκαν στο Google Cloud Storage, μία fine-tuning εργασία (στην πλατφόρμα αποκαλείται ως *tuning job*) μπορούσε να ξεκινήσει χρησιμοποιώντας μία από τις δύο υποστηριζόμενες μεθόδους: μέσω της γραφικής διεπαφής χρήστη (Graphical User Interface ή GUI) του Vertex AI Studio ή μέσω του Python SDK. Και οι δύο προσεγγίσεις απαιτούν τον καθορισμό του βασικού μοντέλου, των URI (Uniform Resource Identifier) των συνόλων δεδομένων, της περιοχής και των προαιρετικών παραμέτρων συντονισμού.

Χρησιμοποιώντας το GUI, η διαδικασία ξεκινά στη διεπαφή του Vertex AI [145]. Αφού επιλεχθεί "Create tuned model", ο χρήστης συμπληρώνει μια φόρμα με τα ακόλουθα πεδία:

- **Βασικό μοντέλο**: επιλέγεται ως `gemini-2.0-flash-lite-001`.
- **Σύνολο δεδομένων εκπαίδευσης**: διεύθυνση URI προς το αρχείο εκπαίδευσης JSONL στο GCS.
- **Σύνολο δεδομένων επικύρωσης**: (προαιρετική) διεύθυνση URI προς το αρχείο επικύρωσης JSONL στο GCS.
- **Περιοχή**: γεωγραφική θέση της εργασίας ρύθμισης (π.χ. `us-central1`).
- **Παράμετροι συντονισμού**: προαιρετική προδιαγραφή του αριθμού εποχών (`epoch_count`), του μεγέθους του προσαρμογέα (`adapter_size`) και του πολλαπλασιαστή του ρυθμού μάθησης (`learning_rate_multiplier`).

Στις Εικόνες 4.2 και 4.3 φαίνονται τα αντίστοιχα πεδία στο προαναφερόμενο GUI.

Μετά την υποβολή, η fine-tuning εργασία μπαίνει σε ουρά και δημιουργείται ένας ειδικός πίνακας ελέγχου. Αυτός ο πίνακας ελέγχου εμφανίζει μετρήσεις προόδου, όπως απώλεια εκπαίδευσης (`training_loss`), ακρίβεια διακριτικών (`token_accuracy`) και αριθμός δειγμάτων (`sample_count`) σε πραγματικό χρόνο. Αυτές οι μετρήσεις ενημερώνονται οπτικά καθ' όλη τη διάρκεια της εργασίας και **δεν** είναι διαθέσιμες μέσω προγραμματιστικών μέσων. Στις Εικόνες 4.4 και 4.5 φαίνονται οι μετρήσεις κατά τη διάρκεια και στο τέλος μίας εργασίας.

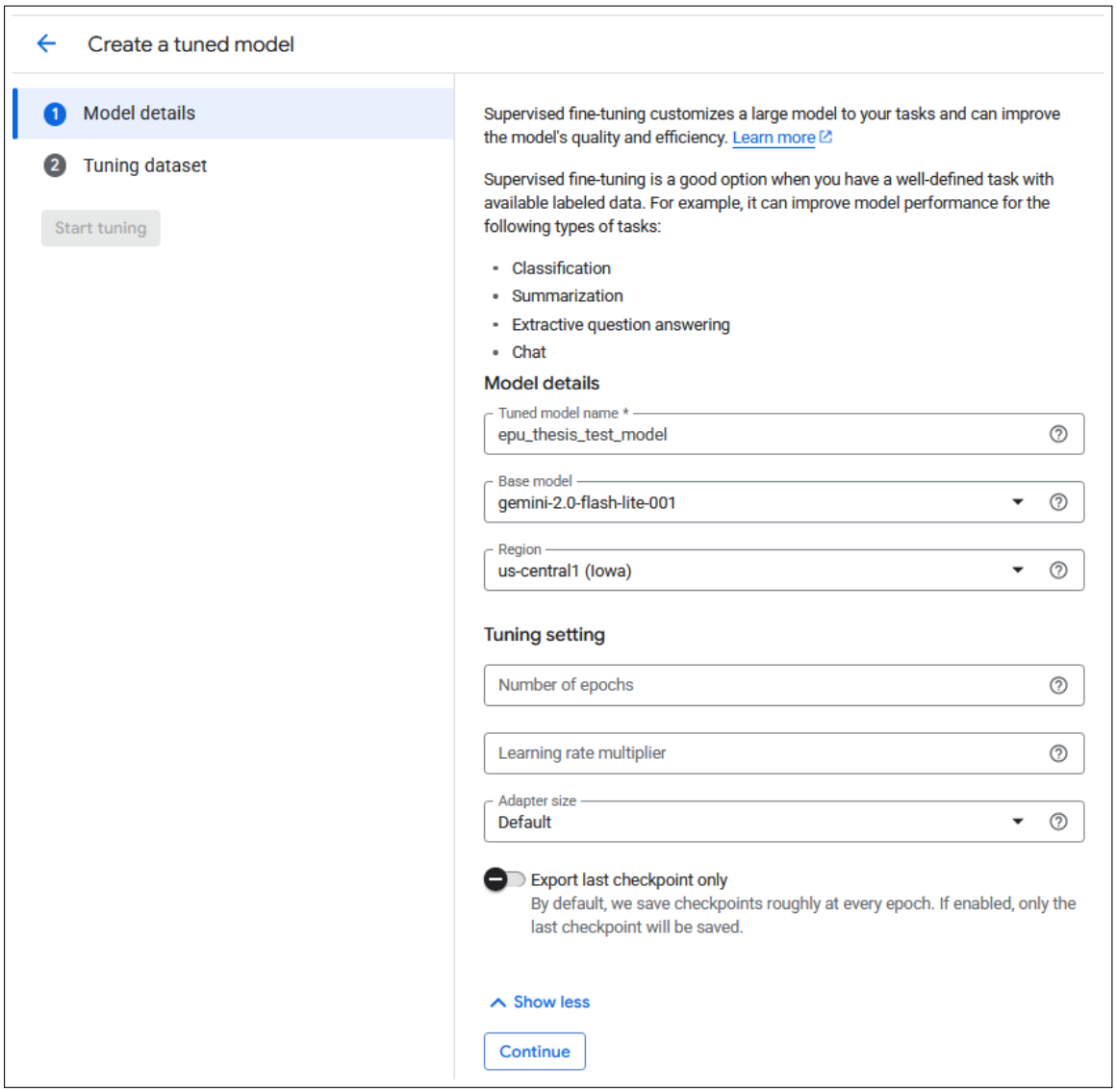
Εναλλακτικά, η ίδια fine-tuning εργασία εργασία μπορεί να εκτελεστεί προγραμματικά χρησιμοποιώντας το Vertex AI Python SDK. Το SDK παρέχει μια διεπαφή λειτουργίας για τον ορισμό και την εκκίνηση μιας εποπτευόμενης εργασίας λεπτού συντονισμού χρησιμοποιώντας τη μέθοδο `sft.train()`. Αν και αυτή η προσέγγιση προσφέρει αυτοματοποίηση και αναπαραγωγιμότητα, εκτυπώνει μόνο την αναφορά της εργασίας και τη διεύθυνση URL παρακολούθησης στο τερματικό. Οι λεπτομερείς μετρήσεις και η πρόοδος πρέπει να εξακολουθούν να είναι προσβάσιμες μέσω του GUI.

Ο κώδικας Python που χρησιμοποιείται για την εκκίνηση, την καταχώριση και τον έλεγχο των fine-tuning εργασιών παρέχεται στο Παράρτημα Γ΄ Python SDK για Vertex AI Tuning.

### 4.3.2 Παράμετροι εκπαίδευσης και έξοδοι μετρήσεων

Το Vertex AI επιτρέπει στους χρήστες να προσαρμόσουν τρεις υπερπαραμέτρους εκπαίδευσης κατά τη διάρκεια του λεπτού συντονισμού:

- **Epoch Count**: αριθμός πλήρων περασμάτων από το σύνολο δεδομένων εκπαίδευσης. Οι υψηλότερες τιμές αυξάνουν γενικά το βάθος μάθησης, αλλά συνεπάγονται μεγαλύτερο κόστος και κίνδυνο υπερπροσαρμογής.
- **Adapter Size**: ελέγχει τη διαστατικότητα των εκπαιδευόμενων επιπέδων που εισάγονται στο παγωμένο βασικό μοντέλο. Οι μεγαλύτερες τιμές επιτρέπουν πιο σύνθετη προσαρμογή, αλλά απαιτούν περισσότερα δεδομένα και υπολογισμούς.
- **Learning Rate Multiplier**: κλιμακώνει τον ρυθμό μάθησης του βελτιστοποιητή. Οι υψηλότερες τιμές επιταχύνουν τη σύγκλιση, αλλά μπορεί να αποσταθεροποιήσουν την εκπαίδευση εάν οριστούν πολύ επιθετικά.



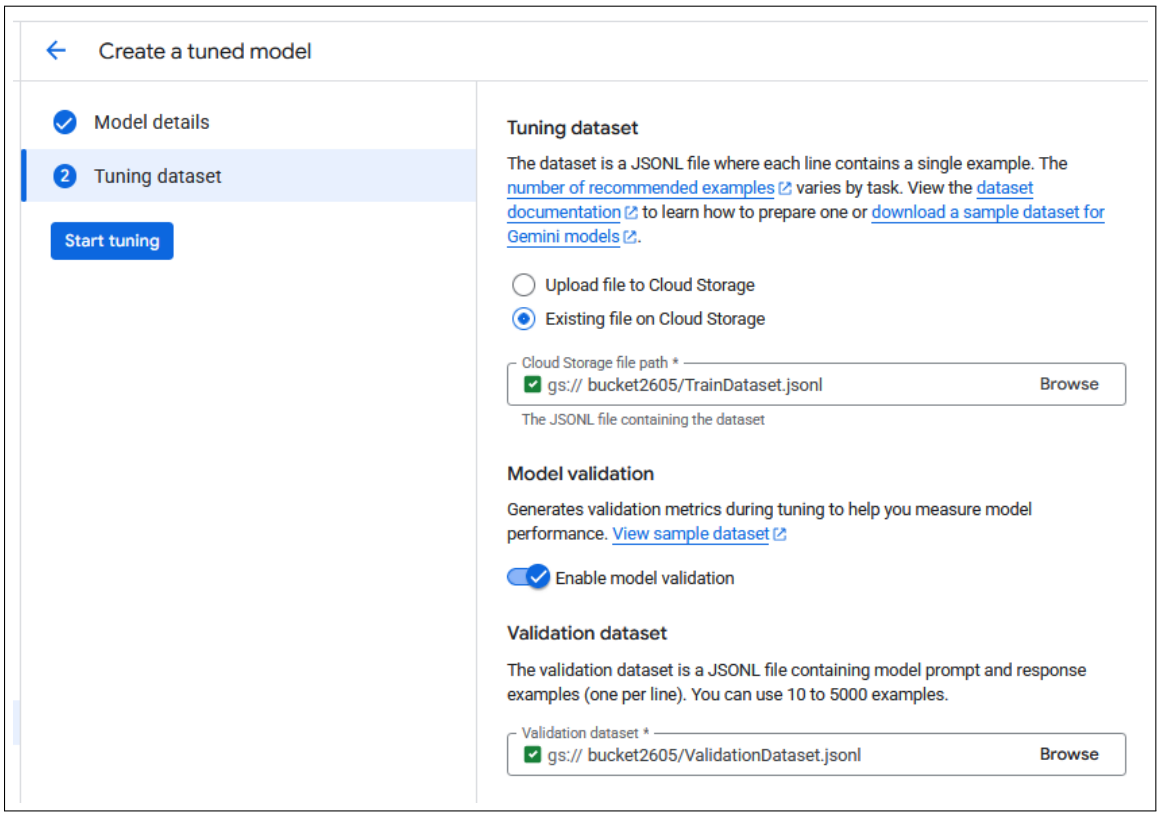

Σχήμα 4.2: *Βήμα 1 Fine-tuning Εργασίας: Επιλογή ονόματος fine-tuned μοντέλου, βασικού μοντέλου και περιοχής. Εάν οι παραμέτροι δεν λάβουν τιμή, τότε αποκτούν την προκαθορισμένη τιμή.*

Αυτές οι παράμετροι μπορούν να παραμείνουν στις προεπιλεγμένες τιμές τους ή να αντικατασταθούν κατά τη ρύθμιση της εργασίας. Οι προεπιλεγμένες τιμές επιλέγονται εσωτερικά από την πλατφόρμα με βάση προηγούμενη ανάλυση απόδοσης.

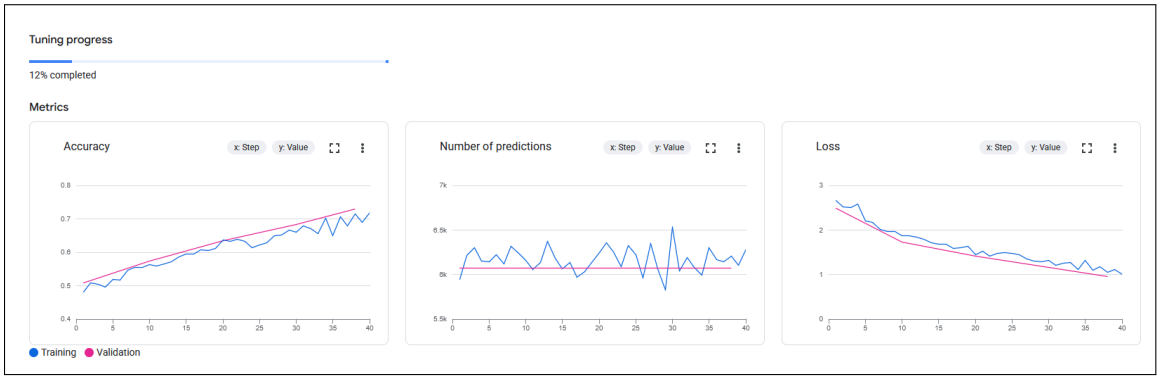
Μόλις ξεκινήσει η εργασία, η πλατφόρμα υπολογίζει διάφορες εσωτερικές μετρήσεις κατά τη διάρκεια των φάσεων εκπαίδευσης και επικύρωσης. Αναφέρονται οι ακόλουθες μετρήσεις:

**Μετρήσεις εκπαίδευσης**

- `/train_total_loss`: συνολική απώλεια στο σύνολο δεδομένων εκπαίδευσης.
- `/train_fraction_of_correct_next_step_preds`: ακρίβεια σε επίπεδο token του μοντέλου κατά τη διάρκεια της εκπαίδευσης.
- `/train_num_predictions`: συνολικός αριθμός προβλεπόμενων token.




Σχήμα 4.3: *Βήμα 2 Fine-tuning Εργασίας: Επιλογή συνόλου δεδομένων εκπαίδευσης και (προαιρετικά) επικύρωσης.*

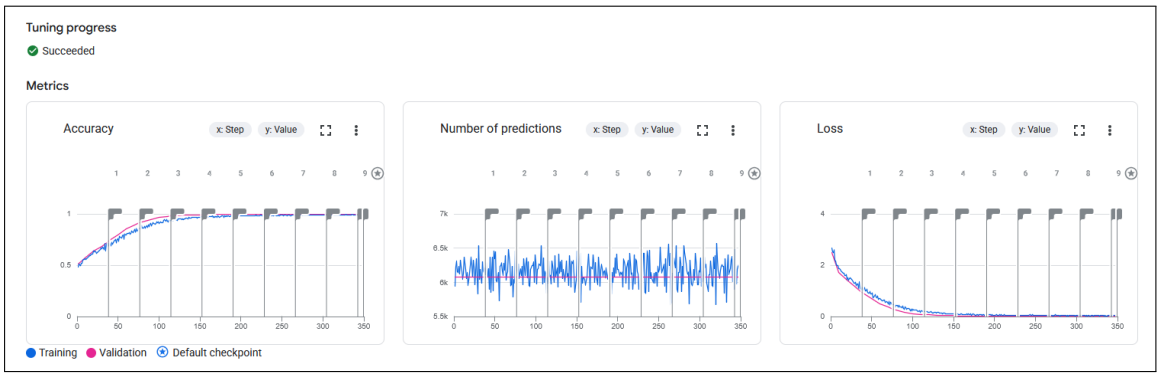



Σχήμα 4.4: *Μετρικές κατά τη διάρκεια μίας tuning εργασίας.*

**Μετρήσεις επικύρωσης**

- `/eval_total_loss`: συνολική απώλεια στο σύνολο δεδομένων επικύρωσης.
- `/eval_fraction_of_correct_next_step_preds`: ακρίβεια σε επίπεδο token στα παραδείγματα επικύρωσης.
- `/eval_num_predictions`: συνολικός αριθμός προβλεπόμενων token στην επικύρωση.

Αυτές οι μετρήσεις είναι ορατές μόνο μέσω της γραφικής διεπαφής και ανανεώνονται περιοδικά κατά την εκτέλεση της εργασίας. Η εμπειρική παρατήρηση έδειξε ότι οι μετρήσεις επικύρωσης δειγματοληπτούνται περίπου μία φορά κάθε τέσσερις εποχές, μια συχνότητα




Σχήμα 4.5: *Μετρικές στο τέλος μίας tuning εργασίας.*

που ενημέρωσε την επακόλουθη ανάλυση κατά τη ρύθμιση των υπερπαραμέτρων.

### 4.3.3 Άπληστη αναζήτηση υπερπαραμέτρων

Για να προσδιοριστεί η βέλτιστη ρύθμιση για τον εικονικό βοηθό, πραγματοποιήσαμε μια άπληστη αναζήτηση υπερπαραμέτρων στις τρεις διαθέσιμες παραμέτρους: αριθμός εποχών (`epoch_count`), μέγεθος προσαρμογέα (`adapter_size`) και πολλαπλασιαστής ρυθμού μάθησης (`learning_rate_multiplier`). Σε κάθε στάδιο, μια παράμετρος μεταβαλλόταν ενώ οι άλλες παρέμεναν σταθερές, επιτρέποντάς μας να αξιολογήσουμε την επίδραση κάθε μεμονωμένης αλλαγής στην απόδοση του μοντέλου. Σε όλες τις περιπτώσεις παρακολουθήθηκαν οι ακόλουθες μετρήσεις: ελάχιστη, μέγιστη, μέση και τελευταία απώλεια επικύρωσης. Ο βασικός δείκτης που καθοδηγούσε αυτή τη διαδικασία ήταν η απώλεια επικύρωσης (`eval_total_loss`) από το σύνολο επικύρωσης, και συγκεκριμένα η τελική του τιμή στο τέλος κάθε εργασίας εκπαίδευσης.

**Σάρωση εποχών**

Σε αυτή τη φάση, το μέγεθος του προσαρμογέα και ο ρυθμός μάθησης καθορίστηκαν στις προεπιλεγμένες τιμές τους (1 και 1, αντίστοιχα). Ο αριθμός εποχών δοκιμάστηκε στις τιμές {10, 25, 50, 75, 100}. Παρατηρήθηκε ότι η απώλεια επικύρωσης βελτιώθηκε καθώς αυξήθηκε ο αριθμός εποχών, αλλά η βελτίωση μειώθηκε μετά από 75 εποχές. Επιπλέον, οι μετρήσεις επικύρωσης δειγματοληπτήθηκαν περίπου μία φορά κάθε τέσσερις εποχές, οδηγώντας σε πιο σταθερές εκτιμήσεις μέσης απώλειας σε υψηλότερες τιμές. Διενεργήθηκε επίσης δοκιμή σε 65 εποχές για να επαληθευτεί η τάση. Η διαμόρφωση με 75 εποχές, `Flash.75`, παρείχε τον καλύτερο συμβιβασμό μεταξύ βάθους μάθησης και αποδοτικότητας. Στον Πίνακα 4.2 φαίνονται αναλυτικά οι τιμές αυτού του βήματος.

**Σάρωση μεγέθους προσαρμογέα**

Στη συνέχεια, καθορίστηκε ο αριθμός εποχών σε 75 και ο ρυθμός μάθησης σε 1, και μεταβλήθηκε το μέγεθος του προσαρμογέα στις επιτρεπόμενες τιμές {1, 4, 8, 16}. Ενώ η μικρότερη απώλεια καταγράφηκε στο μέγεθος προσαρμογέα 8, η διαφορά μεταξύ των μεγεθών προσαρμογέα 8 και 16 ήταν οριακή. Λαμβάνοντας υπόψη τόσο την τελευταία όσο και τη μέση απώλεια επικύρωσης, το μέγεθος προσαρμογέα 8, `Flash.75.8.1`, επιλέχθηκε ως το βέλτιστο. Στον Πίνακα 4.3 φαίνονται αναλυτικά οι τιμές αυτού του βήματος.

| **Epoch** | Samples | Min Val-Loss | Max Val-Loss | Avg Val-Loss | **Last Val-Loss** |
|---|---|---|---|---|---|
| 10 | 4 | 1.51199 | 3.03332 | 2.02750 | 1.51199 |
| 25 | 7 | 0.49142 | 0.86092 | 0.66329 | 0.86092 |
| 50 | 13 | 0.02006 | 3.03332 | 0.82377 | 0.02006 |
| 65 | 17 | 0.014620 | 3.03332 | 0.63367 | 0.01523 |
| **75** | 19 | 0.01263 | 3.03332 | 0.56842 | **0.01550** |
| 100 | 25 | 0.01263 | 3.03332 | 0.43573 | 0.01580 |

Πίνακας 4.2: *Τιμές Απωλειών Επικύρωσης Άπληστης Αναζήτησης για Πλήθος Εποχών*

| **AS** | Samples | Min Val-Loss | Max Val-Loss | Avg Val-Loss | **Last Val-Loss** |
|---|---|---|---|---|---|
| 1 | 19 | 0.01263 | 3.03332 | 0.56842 | 0.01550 |
| 4 | 19 | 0.01217 | 3.03256 | 0.55730 | 0.01227 |
| **8** | 19 | 0.01035 | 3.03307 | 0.55351 | **0.01035** |
| 16 | 19 | 0.01284 | 3.03194 | 0.55121 | 0.01284 |

Πίνακας 4.3: *Τιμές Απωλειών Επικύρωσης Άπληστης Αναζήτησης για Μέγεθος Προσαρμογέα*

#### Σάρωση ρυθμού μάθησης

Τέλος, αξιολογήθηκε ο πολλαπλασιαστής του ρυθμού μάθησης χρησιμοποιώντας τη διαμόρφωση `Flash.75.8.LR`, για τιμές `LR` {0.5, 1, 2, 3}, ακολουθούμενη από μια ενδιάμεση δοκιμή στο 1.5. Ενώ το $LR = 3$ είχε ως αποτέλεσμα τη χαμηλότερη μέση απώλεια, το $LR = 1$ απέδωσε την καλύτερη τελική απώλεια επικύρωσης. Η ενδιάμεση τιμή (1.5) έδειξε πολλά υποσχόμενη απόδοση, αλλά δεν ήταν καλύτερη από το $LR = 1$. Ως αποτέλεσμα, η διαμόρφωση `Flash.75.8.1` επιβεβαιώθηκε ως η **τελική** ρύθμιση για την ανάπτυξη. Στον Πίνακα 4.4 φαίνονται αναλυτικά οι τιμές αυτού του βήματος.

| **LR** | Min Val-Loss | Max Val-Loss | Avg Val-Loss | **Last Val-Loss** |
|---|---|---|---|---|
| 0.5 | 0.01785 | 3.03364 | 0.75949 | 0.01785 |
| **1** | 0.01035 | 3.03307 | 0.55351 | **0.01035** |
| 1.5 | 0.01093 | 3.02987 | 0.47117 | 0.01093 |
| 2 | 0.01040 | 3.02886 | 0.42615 | 0.01341 |
| 3 | 0.01174 | 3.02448 | 0.37185 | 0.01449 |

Πίνακας 4.4: *Τιμές Απωλειών Επικύρωσης Άπληστης Αναζήτησης για Ρυθμό Μάθησης*

#### Σκέψεις σχετικά με το κόστος

Το Vertex AI χρεώνει τις fine-tuning εργασίες με βάση τη διάρκεια υπολογισμού και τη χρήση πόρων, συμπεριλαμβανομένου του μεγέθους του συνόλου δεδομένων και του αριθμού των βημάτων εκπαίδευσης. Κάθε διαμόρφωση μοντέλου που δοκιμάστηκε στην υπερπαραμετρική σάρωση είχε κόστος ανάλογο με τον αριθμό των εποχών και το μέγεθος του προσαρμογέα που χρησιμοποιήθηκε. Ενώ οι εργασίες με μικρότερες διαμορφώσεις (π.χ. 10 εποχές, μέγεθος προσαρμογέα 1) ήταν φθηνές, οι μεγαλύτερες εργασίες —ιδιαίτερα εκείνες που αφορούσαν μεγέθη προσαρμογέα 8 και 16— συνέβαλαν σημαντικά στη συνολική

δαπάνη.

Το πλήρες πείραμα αναζήτησης σε όλες τις εποχές, τα μεγέθη προσαρμογέων και τους ρυθμούς μάθησης είχε ως αποτέλεσμα μια συνολική δαπάνη περίπου 172.40€. Ο αριθμός αυτός αντικατοπτρίζει τόσο το κόστος των εργασιών εκπαίδευσης όσο και τον υπολογιστικό προϋπολογισμό που καταναλώθηκε στο πλαίσιο της δωρεάν κατανομής πιστώσεων που παρέχεται από το Google Cloud. Αυτοί οι περιορισμοί διαμόρφωσαν άμεσα τη δομή της στρατηγικής αναζήτησής της παρούσας εργασίας, ευνοώντας τις άπληστες σαρώσεις ενός παραμέτρου έναντι των εξαντλητικών αναζητήσεων πλέγματος.

Κεφάλαιο 5

# Σχολιασμός Αποτελεσμάτων

Αυτό το κεφάλαιο παρουσιάζει την αξιολόγηση του fine-tuned βοηθού μέσω ενός συνδυασμού θεωρητικών μετρήσεων, χειροκίνητης βαθμολόγησης και πραγματικών δοκιμών χρηστών. Ξεκινώντας με τις βαθμολογίες BLEU και ROUGE σε ένα σύνολο δοκιμών που δεν χρησιμοποιήθηκε στην εκπαίδευση, η ανάλυση εισάγει στη συνέχεια μια ενίσχυση του μοντέλου με αναγνώριση του περιβάλλοντος χρησιμοποιώντας προτροπές βασισμένες σε λέξεις-κλειδιά. Το *ενισχυμένο* αυτό μοντέλο επαναξιολογείται μετά τη βελτίωση, τόσο ποσοτικά όσο και μέσω ανθρώπινης κρίσης. Οι τελικές δοκιμές περιλαμβάνουν την ενσωμάτωση σε μια τοπική έκδοση της πλατφόρμας *ENERGATE* και τη βαθμολόγηση από πραγματικούς χρήστες, προσφέροντας μια ολοκληρωμένη εικόνα της απόδοσης και της πρακτικής βιωσιμότητας του βοηθού.

## 5.1 Αρχική αξιολόγηση μοντέλου

### 5.1.1 Δημιουργία του συνόλου δοκιμών

Για να αξιολογηθεί η απόδοση του βοηθού μετά το fine-tuning, δημιουργήθηκε ένα ειδικό σύνολο δοκιμών αξιοποιώντας μια έβδομη γλωσσική παραλλαγή των ερωτήσεων επικύρωσης. Αυτή η διαδικασία ακολούθησε την ίδια μεθοδολογία που χρησιμοποιήθηκε για τη δημιουργία των συνόλων δεδομένων εκπαίδευσης και επικύρωσης, όπως αυτή περιγράφηκε στην Υποενότητα 4.2.4 Μετατροπή της βάσης γνώσεων σε JSONL, με στόχο την προσομοίωση ρεαλιστικών εισόδων χρηστών που δεν είχαν παρατηρηθεί κατά την εκπαίδευση του μοντέλου.

Συγκεκριμένα, επιλέχθηκαν 60 ερωτήσεις εξάγοντας τις μονές καταχωρήσεις (δηλαδή, προτροπές με δείκτη $2 \cdot n + 1$) από το σύνολο επικύρωσης. Στη συνέχεια, κάθε επιλεγμένη προτροπή αναδιατυπώθηκε σε μια νέα παραλλαγή, διατηρώντας τη σημασιολογική συνοχή με το πρωτότυπο, αλλά εισάγοντας νέα συντακτική δομή.

Το σύνολο δεδομένων που προέκυψε κωδικοποιήθηκε σε μορφή `jsonl`, όπου κάθε γραμμή ακολουθούσε το παρακάτω σχήμα:

```
{
  "id": <ID>,
  "prompt": "<PROMPT>",
  "ideal_response": "<IDEAL_RESPONSE>",
  "model_response": "<MODEL_RESPONSE>"
```

```
}
```

Το id αντιστοιχεί στον αρχικό δείκτη στο σύνολο επικύρωσης (ώστε να γίνει εύκολα η αντιστοιχία των προτροπών μεταξύ των συνόλων δοκιμής και επικύρωσης), το prompt είναι η νέα παραλλαγή, το ideal_response είναι η απάντηση αναφοράς που έχει ήδη χρησιμοποιηθεί ίδια στην εκπαίδευση και την επικύρωση, και το model_response αρχικά παραμένει κενό για να συμπληρωθεί με την έξοδο του μοντέλου.

### 5.1.2 Αλληλεπίδραση μοντέλου και συλλογή απαντήσεων

Το βελτιστοποιημένο μοντέλο Flash.75.8.1 αναπτύχθηκε στο Vertex AI και ερωτήθηκε προγραμματικά χρησιμοποιώντας κώδικα Python για να συμπληρώσει το πεδίο model_response στο σύνολο δεδομένων δοκιμής. Ο κώδικας δημιούργησε αρχικά μια συνεδρία συνομιλίας με το μοντέλο και στη συνέχεια επανέλαβε διαδοχικά κάθε προτροπή, καταγράφοντας την αντίστοιχη απάντηση.

Επιβλήθηκε καθυστέρηση 10 δευτερολέπτων μεταξύ διαδοχικών ερωτήσεων χρησιμοποιώντας τη συνάρτηση time.sleep(10). Αυτό ήταν απαραίτητο για να αποφευχθεί ο περιορισμός της υπηρεσίας ή η προσωρινή απόρριψη αιτήσεων από την πλατφόρμα, η οποία θα μπορούσε να ερμηνεύσει την πρόσβαση υψηλής συχνότητας ως καταχρηστική συμπεριφορά.

Το παρακάτω απόσπασμα κώδικα απεικονίζει τη λογική συλλογής απαντήσεων. Τα credentials (your_project_id) και τα endpoints (your_endpoint_id) έχουν παραλειφθεί για λόγους ασφαλείας:

```
import time
import json
import vertexai
from vertexai.generative_models import GenerativeModel

# Initialize Vertex AI with your project settings.
vertexai.init(
    project="your_project_id",
    location="us-central1",
    api_endpoint="us-central1-aiplatform.googleapis.com"
)

# Create a model instance using your endpoint.
model = GenerativeModel("projects/your_project_id/locations/us-central1/
    endpoints/your_endpoint_id")

input_file = "TestKeywordDataset.jsonl"
output_file = "TestKeyword_Responses.jsonl"

# Load all records from the input file.
records = []
with open(input_file, "r", encoding="utf-8") as fin:
    for line in fin:
        if line.strip():
            records.append(json.loads(line))
```

```
# Process the records.
for record in records:
    prompt = record.get("prompt", "")
    print(f"Processing prompt (id: {record.get('id')}): {prompt}")

    # Start a chat session and send the prompt to the model.
    chat = model.start_chat()
    try:
        response = chat.send_message(prompt)
        record["model_response"] = response.text
        print(f"Model response (id: {record.get('id')}): {response.text}\n"
            )
    except Exception as e:
        print(f"Error processing prompt (id: {record.get('id')}): {e}")

    # Pause for 10 seconds before sending the next prompt.
    time.sleep(10)

# Write all records (both updated and untouched) to the output file.
with open(output_file, "w", encoding="utf-8") as fout:
    for record in records:
        fout.write(json.dumps(record) + "\n")

print(f"Processing complete. Updated records saved in {output_file}")
```

Το αρχείο εξόδου `TestKeyword_Responses.jsonl` περιείχε πλέον το ολοκληρωμένο σύνολο δεδομένων αξιολόγησης με αντιστοιχισμένες απαντήσεις αναφοράς (`ideal_response`) και μοντέλου (`model_response`) για κάθε ερώτηση.

### 5.1.3 Υπολογισμός Μετρήσεων

Αφού καταγράφηκαν οι απαντήσεις του μοντέλου, το επόμενο βήμα περιελάμβανε τον υπολογισμό αυτόματων μετρήσεων ομοιότητας κειμένου για την αξιολόγηση της απόδοσης. Ο κώδικας αξιολόγησης επεξεργάστηκε το αρχείο `TestKeyword_Responses.jsonl` και υπολόγισε τις βαθμολογίες BLEU και ROUGE για κάθε ζεύγος προτροπής-απάντησης.

Όπως αναφέρθηκε και στην Ενότητα 2.3 Μετρικές και σύνολα αξιολόγησης, το BLEU μετρά την επικάλυψη $n$-gram σε επίπεδο token, προσαρμοσμένη με μια λειτουργία εξομάλυνσης για τη διαχείριση μικρότερων ακολουθιών. Το ROUGE υπολογίζει αναδρομικά την επικάλυψη (recall-based overlap), συμπεριλαμβανομένων των ROUGE-1 (unigrams), ROUGE-2 (bigrams) και ROUGE-L (μακρύτερη κοινή υποακολουθία).

Ο κώδικας υπολογισμού των μετρήσεων έχει ως εξής:

```
import json
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
from rouge import Rouge

rouge = Rouge()
bleu_scores = []
rouge_1_scores = []
```

```
rouge_2_scores = []
rouge_l_scores = []
smooth_fn = SmoothingFunction().method1

with open("TestKeyword_Responses.jsonl", "r") as file:
    for line in file:
        data = json.loads(line)
        ideal_tokens = data["ideal_response"].split()
        model_tokens = data["model_response"].split()
        bleu = sentence_bleu([ideal_tokens], model_tokens,
            smoothing_function=smooth_fn)
        bleu_scores.append(bleu)
        scores = rouge.get_scores(data["model_response"], data["
            ideal_response"])
        rouge_1_scores.append(scores[0]['rouge-1']['f'])
        rouge_2_scores.append(scores[0]['rouge-2']['f'])
        rouge_l_scores.append(scores[0]['rouge-l']['f'])

avg_bleu = sum(bleu_scores) / len(bleu_scores)
avg_rouge1 = sum(rouge_1_scores) / len(rouge_1_scores)
avg_rouge2 = sum(rouge_2_scores) / len(rouge_2_scores)
avg_rougeL = sum(rouge_l_scores) / len(rouge_l_scores)

print("Average BLEU Score: {:.4f}".format(avg_bleu))
print("Average ROUGE-1 F-Score: {:.4f}".format(avg_rouge1))
print("Average ROUGE-2 F-Score: {:.4f}".format(avg_rouge2))
print("Average ROUGE-L F-Score: {:.4f}".format(avg_rougeL))
```

Οι τελικές μετρήσεις αξιολόγησης για το μοντέλο `Flash.75.8.1` ήταν:

- Μέση βαθμολογία BLEU: 0.4877
- Μέση βαθμολογία ROUGE-1 F-Score: 0.6201
- Μέση βαθμολογία ROUGE-2 F-Score: 0.5332
- Μέση βαθμολογία ROUGE-L F-Score: 0.6055

Οι βαθμολογίες υποδηλώνουν μέτρια ευθυγράμμιση μεταξύ των αποτελεσμάτων του μοντέλου και των ιδανικών απαντήσεων. Η βαθμολογία *BLEU* 0.4877 αντανακλά αποδεκτή επικάλυψη n-gram, υποδηλώνοντας ότι το μοντέλο παράγει συντακτικά συνεκτικές απαντήσεις, αλλά ενδέχεται να μην αναπαράγει με συνέπεια βασικές φράσεις από την αναφορά. Οι τιμές *ROUGE-1* και *ROUGE-L* άνω του 0.6 δείχνουν ότι το μοντέλο διατηρεί ένα σημαντικό μέρος του σχετικού περιεχομένου και της δομής, κάτι που είναι επιθυμητό για ενημερωτικές απαντήσεις. Η τιμή *ROUGE-2* στο 0.5332 υποδηλώνει επαρκή καταγραφή των μοτίβων bi-gram, αν και υπάρχει περιθώριο βελτίωσης όσον αφορά τη διατήρηση των λεπτομερών λεξικών λεπτομερειών. Σε σύγκριση με τα τυπικά benchmarks, αυτές οι τιμές είναι αποδεκτές, αλλά δεν είναι οι πιο προηγμένες. Δεδομένου ότι το μοντέλο έγινε fine-tuned σε ένα μικρό, domain-specific σύνολο δεδομένων, τα αποτελέσματα θεωρούνται ικανοποιητικά και αποτελούν μια αξιόπιστη βάση για την αξιολόγηση του *ενισχυμένου* μοντέλου που παρουσιάζεται στην επόμενη ενότητα.

## 5.2 Ενίσχυση μοντέλου με αναγνώριση τοποθεσίας

### 5.2.1 Κίνητρο και σχεδιασμός λέξεων-κλειδιών

Ο στόχος της ενίσχυσης των προτροπών με την πληροφορίας της τοποθεσίας είναι να προσομοιώσει μια μορφή εντοπισμού του χρήστη, παρέχοντας στον βοηθό γνώση σχετικά με την τρέχουσα θέση του χρήστη εντός της πλατφόρμας *ENERGATE*. Αυτό επιτυγχάνεται με την προσθήκη λέξεων-κλειδιών (keywords) συγκεκριμένων για την τοποθεσία στις προτροπές του χρήστη. Αυτές οι λέξεις-κλειδιά χρησιμεύουν ως σημασιολογικές βάσεις, που προέρχονται από τη δομή του συστήματος δρομολόγησης URL της πλατφόρμας.

Κάθε λέξη-κλειδί αντιστοιχεί σε μια ξεχωριστή ενότητα της πλατφόρμας, όπως καρτέλες για πληροφορίες κτιρίων, επιλογές χρηματοδότησης, οριστικοποίηση προσφορών, διαχείριση λογαριασμών ή αντιστοίχιση έργων. Για παράδειγμα, η πλοήγηση στην ενότητα `/private_building_technical` της πλατφόρμας θα υποδηλώνει ότι ο χρήστης αλληλεπιδρά με τεχνικές λεπτομέρειες ενός ιδιωτικού κτιρίου.

Αυτή η μέθοδος προσφέρει δύο βασικά πλεονεκτήματα: (i) μειώνει την απαίτηση από τον χρήστη να έχει πλήρης γνώση του πλαισίου της ερώτησής του καθώς και της ακριβής έκφρασης της και (ii) επιτρέπει στον βοηθό να ιεραρχεί και να ερμηνεύει τις ερωτήσεις πιο αποτελεσματικά, ιδίως σε λειτουργικά υπερφορτωμένες ή οπτικά πολύπλοκες ενότητες της διεπαφής.

### 5.2.2 Τροποποίηση του συνόλου δεδομένων με λέξεις-κλειδιά

Τα σύνολα δεδομένων εκπαίδευσης, επικύρωσης και δοκιμής δημιουργήθηκαν ξανά χρησιμοποιώντας το ίδιο γλωσσικό περιεχόμενο με το αρχικό μοντέλο, αλλά με μία βασική τροποποίηση: η προτροπή του χρήστη σε κάθε `jsonl` γραμμή δεδομένων είχε ως πρόθεμα μία ή περισσότερες λέξεις-κλειδιά. Αυτό το πρόθεμα ακολουθούσε την τυποποιημένη μορφή:

```
Keywords: <keyword>, Question: <original prompt>
```

Για παράδειγμα, η αρχική προτροπή:

```
What is the ENERGATE platform?
```

μετατράπηκε σε:

```
Keywords: homepage, Question: What is the ENERGATE platform?
```

Αυτή η τροποποίηση εφαρμόστηκε σε όλες τις παραλλαγές της προτροπής, εξασφαλίζοντας τη συνέπεια του συνόλου εκπαίδευσης. Κάθε προτροπή αντιστοιχήθηκε στις αντίστοιχες λέξεις-κλειδιά με βάση το θεματικό ή λειτουργικό πεδίο εφαρμογής της στο περιβάλλον εργασίας χρήστη του ENERGATE. Η αντιστοίχιση καθορίστηκε με χειροκίνητη αναθεώρηση της δομής και του περιεχομένου της πλατφόρμας.

Σε περιπτώσεις όπου μια προτροπή ήταν ήδη σαφής — π.χ. ερωτήσεις που αναφέρονται σε λογότυπα, εγγραφή χρήστη ή ρυθμίσεις ειδοποιήσεων — η προσθήκη λέξεων-κλειδιών παραλείφθηκε για να αποφευχθεί η εισαγωγή άσχετου ή παραπλανητικού περιεχομένου.

Αυτές οι εξαιρέσεις ήταν σπάνιες και αντιμετωπίστηκαν ρητά κατά την προετοιμασία του νέου συνόλου δεδομένων.

Ο πλήρης κατάλογος των λέξεων-κλειδιών που χρησιμοποιήθηκαν, μαζί με αντιπροσωπευτικά παραδείγματα προτροπών, περιλαμβάνεται στο Παράρτημα Δ΄ Αντιστοίχιση λέξεων-κλειδιών με προτροπές.

Τα ενημερωμένα αρχεία που δημιουργήθηκαν για αυτή τη διαδικασία επαύξησης ακολουθούσαν το ίδιο σχήμα JSONL όπως περιγράφηκε στην Υποενότητα 4.2.4 Μετατροπή της βάσης γνώσεων σε JSONL, με ενημερωμένα παραδείγματα προτροπών που αντανακλούσαν τη μορφή βάσει λέξεων-κλειδιών.

### 5.2.3 Χρήση ενισχυμένου μοντέλου στην τελική εφαρμογή

Ο σχεδιασμός με προσθήκη λέξεων-κλειδιών έχει άμεσες επιπτώσεις στην ανάπτυξη. Στην παραγωγή, ο βοηθός θα ενσωματωθεί ως ένα διαδραστικό εικονίδιο (widget) συνομιλίας στο οποίο οι χρήστες θα μπορούν να έχουν πρόσβαση από οποιαδήποτε σελίδα της πλατφόρμας *ENERGATE*.

Για να διατηρηθεί η συνέπεια με τα δεδομένα εκπαίδευσης, ο βοηθός που έχει αναπτυχθεί θα εισάγει δυναμικά τη σχετική λέξη-κλειδί στην προτροπή πριν την αποστείλει στο μοντέλο. Αυτό θα γίνει με την εξαγωγή της τρέχουσας διαδρομής URL του χρήστη και την αντιστοίχισή της με την αντίστοιχη λέξη-κλειδί.

Η προτροπή που λαμβάνεται από το μοντέλο κατά την εκτέλεση θα αναδιατυπωθεί προγραμματιστικά ως εξής:

```
Keywords: <mapped keyword>, Question: <user's raw input>
```

Αυτή η μετατροπή είναι αόρατη για τον χρήστη και δεν αλλάζει την εισαγωγή του. Αντίθετα, διασφαλίζει ότι ο βοηθός λαμβάνει το ίδιο δομημένο πλαίσιο που έχει εκπαιδευτεί να ερμηνεύει, διατηρώντας τη συνοχή μεταξύ των υποθέσεων κατά την εκπαίδευση και των αλληλεπιδράσεων σε πραγματικό χρόνο.

Αυτός ο σχεδιασμός επιτρέπει στον βοηθό να γενικεύει τη συμπεριφορά του σε όλες τις καταστάσεις πλοήγησης, παραμένοντας ευαίσθητος στο λειτουργικό πλαίσιο του χρήστη.

## 5.3 Αξιολόγηση Βελτιωμένου Μοντέλου

### 5.3.1 Εκπαίδευση και δοκιμή του Flash.75.8.1-k

Χρησιμοποιώντας τα επαυξημένα σύνολα δεδομένων με λέξεις-κλειδιά όπως περιγράφθηκαν στην Ενότητα 5.2 Ενίσχυση μοντέλου με αναγνώριση τοποθεσίας, μια νέα έκδοση του βοηθού έγινε fine-tuned στο Vertex AI. Η διαμόρφωση της εκπαίδευσης παρέμεινε ίδια με την βέλτιστη ρύθμιση που καθορίστηκε στην Ενότητα 5.1 Αρχική αξιολόγηση μοντέλου, δηλαδή `Flash.75.8.1`, χρησιμοποιώντας 75 εποχές, μέγεθος προσαρμογέα 8 και πολλαπλασιαστή ρυθμού μάθησης 1. Το βελτιωμένο μοντέλο αναφέρεται στο εξής ως `Flash.75.8.1-k`, με την κατάληξη "`-k`" να υποδηλώνει την έκδοση με λέξεις-κλειδιά (keywords).

Εφαρμόστηκε η ίδια μεθοδολογία αξιολόγησης. Το βελτιωμένο σύνολο δοκιμών με λέξεις-κλειδιά `Test_Keywords_Dataset.jsonl` χρησιμοποιήθηκε για τη δημιουργία απαντήσεων από το νέο μοντέλο. Κάθε προτροπή περιελάμβανε σαν πρόθεμα τις λέξεις-κλειδιά με την ίδια δομή που χρησιμοποιήθηκε κατά τη διάρκεια του fine-tuning. Τα αποτελέσματα αποθηκεύτηκαν σε ένα αρχείο `TestKeyword_Responses.jsonl`, με απαντήσεις που αντιστοιχούσαν στις επαυξημένες προτροπές.

Στη συνέχεια, το ίδιο σενάριο αξιολόγησης από την Υποενότητα 5.1.3 Υπολογισμός Μετρήσεων επαναχρησιμοποιήθηκε για τον υπολογισμό των μετρήσεων BLEU και ROUGE στα 60 ζεύγη προτροπών-απαντήσεων. Τα αρχικά αποτελέσματα για το βελτιωμένο μοντέλο ήταν τα εξής:

- Μέση βαθμολογία BLEU: 0.5356
- Μέση βαθμολογία ROUGE-1 F: 0.6655
- Μέση βαθμολογία ROUGE-2 F: 0.5813
- Μέση βαθμολογία ROUGE-L F: 0.6534

Αυτές οι τιμές δείχνουν μια μετρήσιμη βελτίωση σε όλες τις διαστάσεις αξιολόγησης σε σχέση με το βασικό μοντέλο `Flash.75.8.1` (Πίνακας 5.1).

| Μετρική | `Flash.75.8.1` | `Flash.75.8.1-k` |
|---|---|---|
| BLEU | 0.4877 | 0.5356 |
| ROUGE-1 | 0.6201 | 0.6655 |
| ROUGE-2 | 0.5332 | 0.5813 |
| ROUGE-L | 0.6055 | 0.6534 |

Πίνακας 5.1: *Συγκριτικός Πίνακας μετρικών BLEU και ROUGE για τα μοντέλα `Flash.75.8.1` και `Flash.75.8.1-k`*

Αυτές οι βελτιώσεις, που παρατηρήθηκαν σε όλες τις παραλλαγές BLEU και ROUGE, επιβεβαιώνουν τον αντίκτυπο της ελάχιστης δομικής ενίσχυσης στην ευθυγράμμιση της συμπεριφοράς του βοηθού με το πλαίσιο του χρήστη σε σύνθετες, πολυλειτουργικές πλατφόρμες όπως η *ENERGATE*.

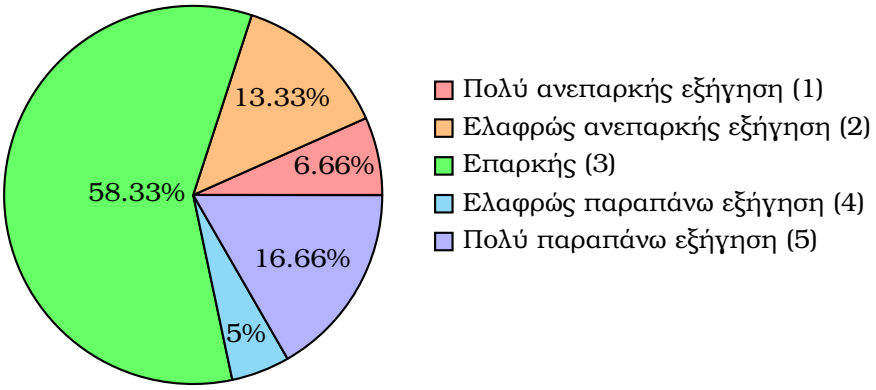
### 5.3.2 Βαθμολόγηση των απαντήσεων του μοντέλου

Εκτός από τις αυτοματοποιημένες αξιολογήσεις μετρικών, πραγματοποιήθηκε μια υποκειμενική αξιολόγηση για την ποιοτική ανάλυση των απαντήσεων του μοντέλου `Flash.75.8.1-k`. Κάθε μία από τις 60 απαντήσεις του μοντέλου στο σύνολο δοκιμών εξετάστηκε και βαθμολογήθηκε σε μια διακριτή κλίμακα *Likert* από 1 έως 5, με βάση την ευθυγράμμισή της με την ιδανική απάντηση αναφοράς (`ideal_response`). Τα κριτήρια αξιολόγησης ορίστηκαν ως εξής:

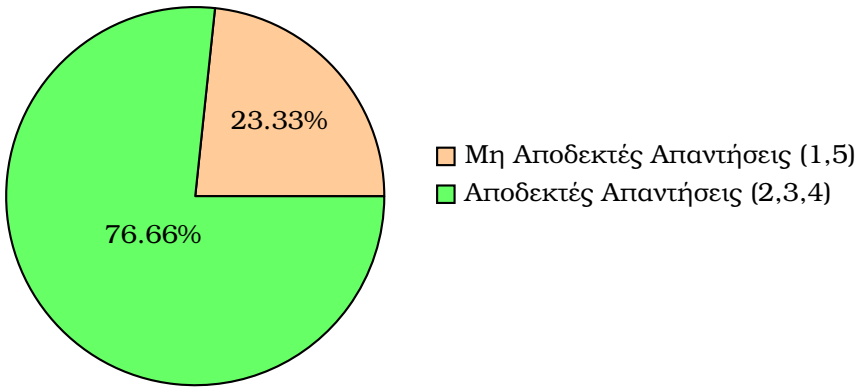
- **1 – Πολύ ανεπαρκής εξήγηση:** Η απάντηση ήταν σαφώς λεκτικά ανεπαρκής, εκτός θέματος ή δεν απαντούσε στην ερώτηση.
- **2 – Ελαφρώς ανεπαρκής εξήγηση:** Η απάντηση ήταν εν μέρει σχετική, αλλά έλειπαν βασικές λεπτομέρειες.

- **3 – Επαρκής / Πανομοιότυπη:** Η απάντηση ήταν ισοδύναμη με την ιδανική αναφορά ως προς το περιεχόμενο και το πεδίο εφαρμογής.
- **4 – Ελαφρώς παραπάνω εξήγηση:** Η απάντηση ήταν σωστή, αλλά περιείχε ελαφρώς παραπάνω ή περιττές λεπτομέρειες.
- **5 – Πολύ παραπάνω εξήγηση:** Η απάντηση περιείχε περιττές λεπτομέρειες που αμβλύνουν τη συνάφεια ή παρεκκλίνουν από το θέμα.

Για να απεικονιστεί η κατανομή, δημιουργήθηκαν δύο διαγράμματα. Το πρώτο απεικονίζει τον αριθμό ανά κατηγορία αξιολόγησης όπως παρουσιάστηκε παραπάνω (Σχήμα 5.1). Το δεύτερο συγκεντρώνει τις αξιολογήσεις (2), (3) και (4) ως *αποδεκτές απαντήσεις* και τις αξιολογήσεις (1) και (5) ως ***μη*** *αποδεκτές απαντήσεις* (δηλαδή εκτός πεδίου εφαρμογής, Out of Scope ή OoS), αντιπροσωπεύοντας μια πιο χαλαρή ταξινόμηση της επιτυχίας (Σχήμα 5.2).




Σχήμα 5.1: *Κατανομή Βαθμολόγησης 60 ερωτημάτων (διακριτές κατηγορίες)*




Σχήμα 5.2: *Κατανομή Βαθμολόγησης 60 ερωτημάτων (ομαδοποιημένες κατηγορίες)*

Οι κατανομές επιβεβαιώνουν ότι η συντριπτική πλειοψηφία των απαντήσεων (76.6%) εμπίπτει στα αποδεκτά όρια ερμηνείας (βαθμολογία 2–4), με το 58.3% να βαθμολογείται ως πλήρως ευθυγραμμισμένο με το ιδανικό (βαθμολογία 3). Ένα μικρότερο ποσοστό (23.4%) θεωρήθηκε εκτός πεδίου εφαρμογής ή μη ευθυγραμμισμένο, είτε λόγω ανεπάρκειας είτε λόγω υπερβολικής επέκτασης.

Αυτά τα ευρήματα συμπληρώνουν τα αποτελέσματα των αυτόματων μετρήσεων, παρέχοντας ανθρώπινα ερμηνεύσιμες πληροφορίες σχετικά με την ποιοτική ορθότητα, τη συνάφεια του περιεχομένου και την πιθανή υπερπροσαρμογή ή παραισθήσεις.

## 5.4 Ανάπτυξη και αξιολόγηση πραγματικούς χρήστες

### 5.4.1 Ενσωμάτωση του chatbot στην πλατφόρμα *ENERGATE*

Για να αποδειχθεί η πρακτική ενσωμάτωση, το μοντέλο `Flash.75.8.1-k` εφαρμόστηκε σε μια τοπική ανάπτυξη της πλατφόρμας *ENERGATE*. Για το μέρος της διεπαφής frontend υλοποιήθηκε σε React-JS ένα κομμάτι κώδικα (`Chatbox.js`) ώστε το μοντέλο να λειτουργεί ως διαδραστικό widget chatbot προσβάσιμο σε όλες τις σελίδες. Το widget ανακτούσε δυναμικά την τρέχουσα θέση URL και εισήγαγε την κατάλληλη λέξη-κλειδί στο πεδίο προτροπής πριν στείλει το ερώτημα στο backend.

Από την πλευρά του διακομιστή, το endpoint διαχειρίστηκε μέσω του Django REST framework. Ο κώδικας `ai_utils.py` περιείχε την βασική λογική για την αλληλεπίδραση με το endpoint στο Vertex AI που φιλοξενούσε το fine-tuned μοντέλο, ενώ το `chat_views.py` παρείχε τη διεπαφή HTTP.

**Frontend (Chatbox.js)**

Ο κώδικας εμφανίζει ένα widget συνομιλίας που μπορεί να μεγενθυθεί σε ένα πλήρες παράθυρο συνομιλίας όταν ενεργοποιηθεί. Παρακολουθεί την τρέχουσα καρτέλα (`curLocation`) και την ενσωματώνει στην προτροπή του χρήστη με τη μορφή:

```
Keywords: <curLocation>. Question: <prompt>
```

Διατηρεί την κατάσταση της συνομιλίας, χειρίζεται τις εισόδους και αποδίδει απαντήσεις σε πραγματικό χρόνο από το backend. Το ιστορικό μηνυμάτων κυλά αυτόματα για να διατηρείται ορατή η τελευταία αλληλεπίδραση.

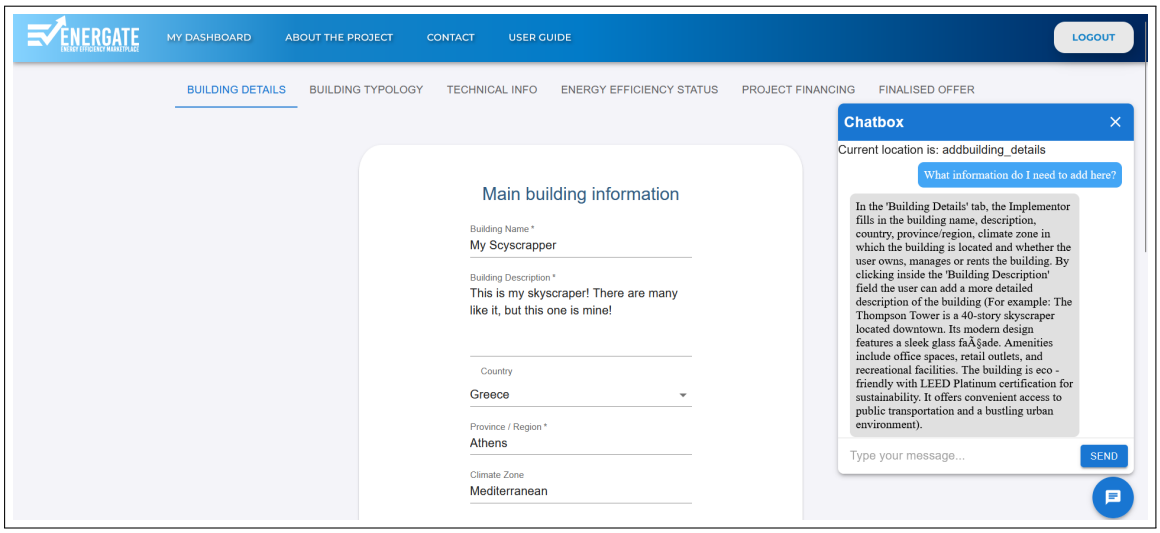
**Backend (ai_utils.py και chat_views.py)**

Το backend αρχικοποιεί τη διεπαφή του μοντέλου μέσω του Vertex AI SDK. Κάθε προτροπή χρήστη που λαμβάνεται μέσω του endpoint `/chat/get_response/` μεταβιβάζεται στο μοντέλο Gemini, η απάντηση του οποίου επιστρέφεται στο frontend.

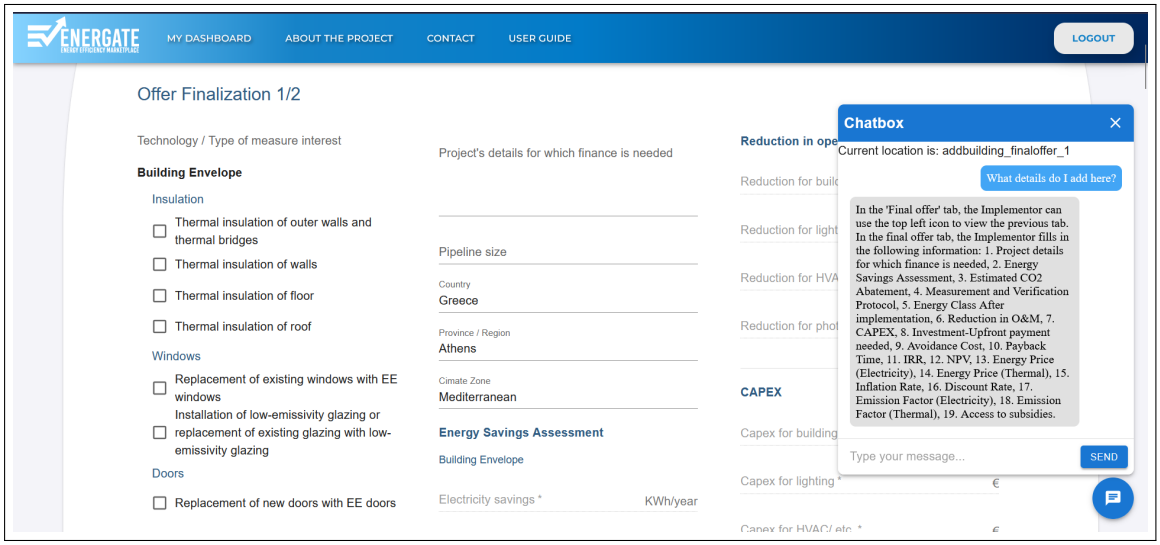
Τα Σχήματα 5.3 και 5.4 παρουσιάζουν το chatbot σε λειτουργία σε δύο διαφορετικές σελίδες της *ENERGATE*. Σε κάθε περίπτωση, η λέξη-κλειδί που χρησιμοποιείται στην επαυξημένη προτροπή αντανακλά το τρέχον πλαίσιο πλοήγησης του χρήστη. Πάνω στο chatbot, φαίνεται πως αλλάζουν τα keywords που δέχεται αναλόγως την τοποθεσία, ενώ η ερώτηση έχει τεθεί εσκεμμένα αόριστη (*Τι κάνω εδώ;*) ώστε να φανεί ξεκάθαρα πως το μοντέλο καταλαβαίνει από το πλαίσιο.

### 5.4.2 Ανθρώπινη αξιολόγηση της απόδοσης του βοηθού

Για να συμπληρωθούν οι αυτοματοποιημένες και χειροκίνητες αξιολογήσεις, πραγματοποιήθηκε μια δοκιμή με πραγματικούς χρήστες. Μια σειρά από 30 μοναδικές ερωτήσεις χρηστών (ηλεκτρολόγοι μηχανικοί, που αλληλεπιδρούσαν για πρώτη φορά με την πλατ-




Σχήμα 5.3: *Chatbot ενεργό στην πρώτη καρτέλα προσθήκης ενός κτηρίου, Building Details*



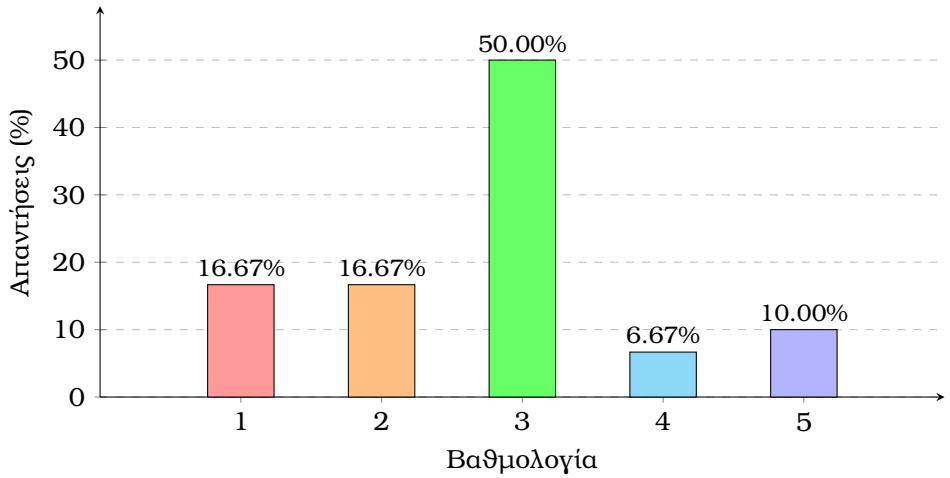

Σχήμα 5.4: *Chatbot ενεργό στην καρτέλα τελικής προσφοράς, Finalised Offer*

φόρμα) υποβλήθηκαν στον βοηθό στην τοπικά εγκατεστημένη εφαρμογή *ENERGATE*. Κάθε χρήστης βαθμολόγησε την απάντηση του βοηθού στην ίδια κλίμακα 1–5 που περιγράφεται στην Υποενότητα 5.3.2 Βαθμολόγηση των απαντήσεων του μοντέλου. Η κατανομή των βαθμολογιών των χρηστών φαίνονται στο Σχήμα 5.5

Αυτά τα αποτελέσματα επικυρώνουν τα ευρήματα των προηγούμενων αξιολογήσεων, με 22 από τις 30 απαντήσεις (73.33%) να βαθμολογούνται εντός του αποδεκτού εύρους (βαθμολογίες 2–4), ενισχύοντας την αξιοπιστία του βοηθού σε ρεαλιστικές συνθήκες χρήσης.

### 5.4.3 Σύνοψη Αποτελεσμάτων

Σε αυτό το κεφάλαιο παρουσιάστηκε η αξιολόγηση του βελτιστοποιημένου βοηθού, ξεκινώντας με αυτοματοποιημένες μετρήσεις (BLEU, ROUGE) και συνεχίζοντας με χειροκίνητη αναθεώρηση και βελτίωση του συνόλου εκπαίδευσης μέσω της εισαγωγής λέξεων-κλειδιών.




Σχήμα 5.5: *Κατανομή Βαθμολόγησης σε 30 ερωτήματα πραγματικών χρηστών*

Η έκδοση με λέξεις-κλειδιά (`Flash.75.8.1-k`) έδειξε σταθερή βελτίωση στην ποιότητα της παραγωγής απαντήσεων. Οι βελτιώσεις μετά την επεξεργασία ενίσχυσαν περαιτέρω την ακρίβεια της βαθμολόγησης.

Το μοντέλο χρησιμοποιήθηκε στη συνέχεια ως βοηθός και ενσωματώθηκε με επιτυχία στην πλατφόρμα *ENERGATE* με δυναμική αναγνώριση του περιβάλλοντος, ενώ οι δοκιμές αλληλεπίδρασης επιβεβαίωσαν την προσαρμοστικότητά του στην τοποθεσία του χρήστη. Τέλος, η εξωτερική αξιολόγηση από πραγματικούς χρήστες επιβεβαίωσε τη συνάφεια, την πληροφόρηση και τη σημασιολογική ευθυγράμμιση σε ένα ευρύ φάσμα ερωτημάτων.

Τα ευρήματα επικυρώνουν την αποτελεσματικότητα του εξειδικευμένου fine-tuning, της ενίσχυσης δομημένων δεδομένων (structured data augmentation) και της προετοιμασίας κατάλληλων προτροπών για την παροχή ισχυρής, ευαίσθητης στο περιβάλλον εικονικής βοήθειας.

# Μέρος III

# Επίλογος

Κεφάλαιο 6

# Επίλογος

Η παρούσα διπλωματική εργασία ασχολήθηκε με ένα κρίσιμο κενό στη χρηστικότητα της πλατφόρμας *ENERGATE*, μίας υπηρεσίας που υποστηρίζεται από την ΕΕ για τη διευκόλυνση των επενδύσεων σε ενεργειακή απόδοση κτιρίων. Η πλατφόρμα παρέμενε απρόσιτη σε πολλούς χρήστες λόγω της πολυπλοκότητάς της, της μη τυποποιημένης ορολογίας και της δημογραφικής ποικιλομορφίας όσον αφορά την ψηφιακή ικανότητα των χρηστών της. Το βασικό πρόβλημα ήταν η ανάγκη να γεφυρωθεί αυτό το εμπόδιο στη χρήση της χωρίς να υπάρχει εξάρτηση από την ανθρώπινη υποστήριξη.

Η προτεινόμενη λύση ήταν ο σχεδιασμός και η ανάπτυξη ενός εικονικού βοηθού βασισμένου σε ένα μεγάλο γλωσσικό μοντέλο, με στόχο να καθοδηγεί τους χρήστες στις λειτουργίες της πλατφόρμας. Μετά από μια αρχική αξιολόγηση διαφόρων LLM, η οποία έγινε με βάση την αρχιτεκτονική, την απόδοση, την αδειοδότηση και τους περιορισμούς υλοποίησης, μια τοπική προσπάθεια υλοποίησης με χρήση LLaMA απορρίφθηκε λόγω περιορισμών υλικού. Στη συνέχεια, επιλέχθηκαν μοντέλα Gemini μέσω της υπηρεσίας Vertex AI της Google Cloud για fine-tuning με χρήση εποπτευόμενης μάθησης.

Ο βοηθός που προέκυψε, βελτιστοποιημένος σε δομημένα ζεύγη ερωτήσεων-απαντήσεων με λέξεις-κλειδιά, που προέρχονταν από τις συχνές ερωτήσεις αλλά και τους οδηγούς χρηστών της *ENERGATE*, μπορούσε να προσαρμόσει τις απαντήσεις του σε σχέση με την τοποθεσία του χρήστη στην πλατφόρμα. Ο βοηθός ενσωματώθηκε με επιτυχία σε μια τοπική παρουσία του frontend της *ENERGATE* και αξιολογήθηκε μέσω βαθμολογιών BLEU/ROUGE, κρίσης εμπειρογνωμόνων και αλληλεπιδράσεων με πραγματικούς χρήστες.

## 6.1 Ακαδημαϊκή συμβολή

Από μεθοδολογική άποψη, η παρούσα διατριβή παρουσιάζει μία ευκόλως αναπαραγώμενη διαδικασία για το fine-tuning ενός LLM για συγκεκριμένους τομείς χρησιμοποιώντας περιορισμένους πόρους. Δείχνει ότι τα προσεκτικά δομημένα σύνολα δεδομένων, που έχουν εμπλουτιστεί συνθετικά μέσω της διαφορετικών συντακτικά προτροπών και της επισήμανσης του πλαισίου μπορούν να αποφέρουν συγκρίσιμα αποτελέσματα ακόμη και σε περιβάλλοντα με περιορισμένους πόρους.

Από τεχνική άποψη, η εργασία αποδεικνύει τη βιωσιμότητα του εποπτευόμενου fine-tuning ειδικού για κάθε πλατφόρμα, χρησιμοποιώντας μία σύγχρονη διαδικασία βασισμένη στο cloud. Η ενσωμάτωση προτροπών ευαίσθητων στην τοποθεσία (μέσω λέξεων-κλειδιών)

βελτίωσε την ακρίβεια των απαντήσεων, επικυρώνοντας έναν μηχανισμό χαμηλού κόστους για τη βελτίωση του μοντέλου, χωρίς να απαιτούνται αρχιτεκτονικές αλλαγές.

Εμπειρικά, ο βοηθός πέτυχε μετρήσιμες βελτιώσεις σε όλες τις τυπικές μετρήσεις (BLEU, ROUGE) και τις ποιοτικές αξιολογήσεις. Το μοντέλο με επαυξημένο σύνολο εκπαίδευσης μέσω λέξεων-κλειδιών (`Flash.75.8.1-k`) υπερείχε του αντίστοιχου βασικού μοντέλου και έλαβε θετική αξιολόγηση τόσο από ειδικούς όσο και από πραγματικούς χρήστες.

Δομικά, η εργασία αυτή συμβάλλει με ένα αρχικό πρωτότυπο για την ενσωμάτωση συστημάτων υποστήριξης βασισμένων σε LLM σε σύνθετες πλατφόρμες, παρέχοντας ένα πρότυπο για μελλοντική έρευνα στον τομέα της ενίσχυσης της χρηστικότητας και της συνεργασίας μεταξύ ανθρώπων και τεχνητής νοημοσύνης σε ψηφιακά ενεργειακά συστήματα.

## 6.2 Περιορισμοί και μελλοντική εργασία

Αν και τα αποτελέσματα επιβεβαιώνουν τη λειτουργική αποτελεσματικότητα του βοηθού, παραμένουν αρκετοί περιορισμοί. Το σύνολο δεδομένων εκπαίδευσης δημιουργήθηκε από ένα σταθερό σύνολο οδηγών χρήστη και συχνών ερωτήσεων, πράγμα που σημαίνει ότι η γλωσσική ποικιλομορφία του πραγματικού κόσμου και οι ακραίες περιπτώσεις δεν εκπροσωπήθηκαν επαρκώς. Επιπλέον, παρά την άμεση ενίσχυση, η κατανόηση του βοηθού παραμένει περιορισμένη στη στατική βάση γνώσεων που υπάρχει κατά τη στιγμή του fine-tuning.

Μελλοντικές εργασίες θα πρέπει να αντιμετωπίσουν αυτούς τους περιορισμούς ενσωματώνοντας ένα δυναμικό σύστημα ανάκτησης γνώσεων (RAG) για να επιτρέψουν την πρόσβαση σε πραγματικό χρόνο σε ενημερωμένο περιεχόμενο της πλατφόρμας. Η επέκταση του συνόλου δεδομένων με αρχεία καταγραφής από πραγματικές αλληλεπιδράσεις χρηστών θα ενίσχυε τον ρεαλισμό και την κάλυψη, ιδίως για πολυγλωσσικές εισόδους και σύνθετες ροές εργασίας. Επιπλέον, το σύνολο δεδομένων μπορεί να επεκταθεί μέσω μιας πιο συστηματικής και λεπτομερούς χαρτογράφησης όλων των λειτουργικών περιοχών της πλατφόρμας, διασφαλίζοντας ότι η κάλυψη του βοηθού αντανακλά το πλήρες εύρος των αλληλεπιδράσεων των χρηστών. Οι βρόχοι ζωντανής ανατροφοδότησης, όπου οι αποτυχημένες ερωτήσεις σχολιάζονται και επανενσωματώνονται, θα μπορούσαν επίσης να βελτιώσουν την ανθεκτικότητα και την ακρίβεια με την πάροδο του χρόνου.

Από τεχνική άποψη, η εισαγωγή ανακατεύθυνσης (fallback) για ασαφείς ή μη υποστηριζόμενες ερωτήσεις, καθώς και η βαθμολόγηση των απαντήσεων με βάση την αξιοπιστία, θα αυξήσουν την εμπιστοσύνη των χρηστών και θα μειώσουν τους κινδύνους παραισθήσεων. Από την άποψη της υλοποίησης σαν εμπορικό προϊόν, η μετάβαση του βοηθού από το πρωτότυπο στην παραγωγή απαιτεί υποδομή παρακολούθησης, έλεγχο πρόσβασης και μηχανισμούς ενημέρωσης που να είναι συμβατοί με την πλατφόρμα.

## 6.3 Τελικές παρατηρήσεις

Η παρούσα διατριβή παρουσιάζει την υλοποίηση ενός πλήρους κύκλου ενός εξειδικευμένου εικονικού βοηθού βασισμένου σε LLM, ο οποίος έχει σχεδιαστεί για να λειτουργεί σε μια δομημένη, ειδική για τον τομέα ψηφιακή πλατφόρμα. Η ανάπτυξη, η ρύθμιση και

η αξιολόγηση του βοηθού επιβεβαιώνουν ότι τα βελτιστοποιημένα γλωσσικά μοντέλα, όταν ενισχύονται με προτροπές με γνώση του περιβάλλοντος, μπορούν να βελτιώσουν σημαντικά την προσβασιμότητα και τη χρηστικότητα σύνθετων συστημάτων, όπως η *ENERGATE*.

Πέρα από την άμεση εφαρμογή του, το έργο χρησιμεύει ως ένα αρθρωτό πλαίσιο για την ενσωμάτωση της εικονικής βοήθειας σε άλλες πλατφόρμες του τομέα της ενέργειας ή σε παρόμοια ρυθμιζόμενα περιβάλλοντα. Γεφυρώνει τις τεχνικές εξελίξεις στα LLM με τις πρακτικές ανάγκες στη διακυβέρνηση της βιωσιμότητας, υποστηρίζοντας μια πιο περιεκτική και αποτελεσματική αλληλεπίδραση μεταξύ των ενδιαφερομένων μερών. Ο βοηθός δεν αποτελεί τελικό προϊόν, αλλά πρωτότυπο για επεκτάσιμα, προσαρμοστικά συστήματα υποστήριξης που βασίζονται σε τεχνητή νοημοσύνη.

# Παραρτήματα

# Παράρτημα Α΄

# Συμπληρωματικές Τεχνικές Έννοιες

Αυτό το παράρτημα περιέχει εκτεταμένο τεχνικό υλικό που αναφέρεται στο Κεφάλαιο 2: Θεωρητικό υπόβαθρο, συμπεριλαμβανομένων των βασικών μηχανισμών της βαθιάς μάθησης, της επεξεργασίας φυσικής γλώσσας και των αρχιτεκτονικών μετασχηματιστών. Παρέχει τις μαθηματικές και αλγοριθμικές διατυπώσεις που υποστηρίζουν τις έννοιες που συζητούνται στο κύριο κείμενο, αλλά εξαιρέθηκαν για λόγους συντομίας.

## Α.1 Βελτιστοποίηση και Γενίκευση

Τα συστήματα βαθιάς μάθησης εκπαιδεύονται με τη μέθοδο της στοχαστικής κατάβασης κλίσης και τις παραλλαγές της, προκειμένου να ελαχιστοποιήσουν την εμπειρική απώλεια σε μεγάλα σύνολα δεδομένων.

Η *Στοχαστική Κατάβαση Κλίσης* είναι ένας επαναληπτικός αλγόριθμος βελτιστοποίησης που χρησιμοποιείται για την ελαχιστοποίηση της συνάρτησης απώλειας στη βαθιά μάθηση. Ενημερώνει τις παραμέτρους του μοντέλου υπολογίζοντας την κλίση της απώλειας σε σχέση με τις παραμέτρους χρησιμοποιώντας ένα τυχαία επιλεγμένο υποσύνολο (μίνι-παρτίδα ή minibatch) των δεδομένων εκπαίδευσης, εισάγοντας θόρυβο που μπορεί να βοηθήσει στην αποφυγή τοπικών ελαχίστων και κρίσιμων σημείων (saddle points) [151]. Ο κανόνας ενημέρωσης είναι:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t; x_i, y_i)$$

όπου $\theta$ είναι οι παράμετροι, $\eta$ είναι ο ρυθμός μάθησης και $\mathcal{L}$ είναι η συνάρτηση απώλειας που αξιολογείται στο δείγμα $(x_i, y_i)$. Έχει ταχύτερες ενημερώσεις από την κατάβαση κλίσης πλήρους παρτίδας (full-batch gradient descent), επιτρέποντας την κλιμάκωση σε μεγάλα σύνολα δεδομένων. Ωστόσο, η σύγκλιση στον αριθμό βημάτων είναι πιο αργή και πιο θορυβώδης από την κατάβαση κλίσης πλήρους παρτίδας[1], αλλά πιο πρακτική σε χώρους υψηλής διάστασης. Το SGD χρησιμοποιείται συχνά με παραλλαγές όπως momentum, RMSProp ή Adam για τη βελτίωση της σταθερότητας και της ταχύτητας σύγκλισης [151].

[1] Η SGD απαιτεί συνήθως πολύ περισσότερα βήματα ενημέρωσης από την κατάβαση κλίσης πλήρους παρτίδας (full-batch gradient descent) για να επιτύχει συγκρίσιμη απώλεια, επειδή κάθε βήμα χρησιμοποιεί μόνο μια μίνι παρτίδα. Ωστόσο, κάθε βήμα SGD επεξεργάζεται πολύ λιγότερα παραδείγματα και μπορεί να εκτελεστεί παράλληλα σε σύγχρονο υλικό. Επομένως, ο συνολικός χρόνος που απαιτείται για την επίτευξη αυτής της απώλειας είναι συνήθως μικρότερος, παρά το μεγαλύτερο πλήθος βημάτων.

Τα υπερπαραμετροποιημένα δίκτυα γενικεύονται όταν ρυθμίζονται σωστά. Οι κλασικές στρατηγικές περιλαμβάνουν την *αποσύνθεση βάρους*, την *πρόωρη διακοπή* και η *τεχνική αποκοπής νευρώνων*. Η *αποσύνθεση βάρους* (weight decay) προσθέτει μια ποινή $L_2$ στον στόχο, συρρικνώνοντας τις παραμέτρους προς το μηδέν και ελέγχοντας την ικανότητα του μοντέλου [152]. Η *πρόωρη διακοπή* (early stopping) σταματά την εκπαίδευση όταν η απώλεια επικύρωσης σταματά να μειώνεται, αποτρέποντας την υπερπροσαρμογή με τον περιορισμό της αποτελεσματικής χωρητικότητας [153]. Η *τεχνική αποκοπής νευρώνων* (dropout) απενεργοποιεί τυχαία τις ενεργοποιήσεις κατά τη διάρκεια της εκπαίδευσης, διακόπτοντας την συν-προσαρμογή (co-adaptation) των ανιχνευτών χαρακτηριστικών (feature detectors) και λειτουργώντας ως μέσος όρος του μοντέλου κατά τη διάρκεια της δοκιμής [154]. Η συμπεριφορά γενίκευσης τέτοιων κανονικοποιητών υποδεικνύει εμπειρικές σχέσεις με τη θεωρία Vapnik–Chervonenkis (VC), η οποία περιορίζει το σφάλμα σε όρους πολυπλοκότητας της κατηγορίας υποθέσεων [155].

### Α.1.1 Deep Learning και η εποχή των transformers

Οι καινοτομίες στην μάθηση πολυεπίπεδης αναπαράστασης κορυφώθηκαν με την έρευνα των LeCun, Bengio και Hinton, που πλαισίωσε τη «βαθιά μάθηση» (deep learning) ως κλιμακωτή ιεραρχική εκμάθηση αναπαράστασης. Η *μάθηση πολυεπίπεδης αναπαράστασης* (multilayer representation learning) στοιβάζει πολλά μη γραμμικά επίπεδα, έτσι ώστε τα υψηλότερα επίπεδα να καταγράφουν όλο και πιο αφηρημένες έννοιες [156]. Τα παλαιότερα μοντέλα ακολουθιών βασίζονταν στην αναδρομή, εφαρμόζοντας επανειλημμένα μια ενημέρωση κρυφής κατάστασης για να επεξεργαστούν διαδοχικά τα tokens [157].

Η αρχιτεκτονική *μετασχηματιστή* (transformer) αντικατέστησε την αναδρομή με την αυτοπροσοχή (self-attention), επιτυγχάνοντας μοντελοποίηση ακολουθιών και επιτρέποντας μαζική παράλληλη εκπαίδευση [2]. Εμπειρικοί «νόμοι κλιμάκωσης» αποκάλυψαν προβλέψιμες βελτιώσεις του νόμου-ισχύος[2] στην απώλεια ως συνάρτηση του μεγέθους του μοντέλου, του μεγέθους του συνόλου δεδομένων και της υπολογιστικής ισχύος [37] – αυτές οι καμπύλες ενέπνευσαν τα σημερινά LLM με δισεκατομμύρια παραμέτρους.

## Α.2 Επεξεργασία Φυσικής Γλώσσας

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing ή NLP) μελετά αλγόριθμους που αναλύουν, παράγουν ή μετασχηματίζουν την ανθρώπινη γλώσσα [159]. Τα κλασικά συστήματα αναλύουν την εργασία σε modular στάδια, καθένα από τα οποία εφαρμόζει όλο και πιο πλούσιους γλωσσικούς φορμαλισμούς στο ακατέργαστο κείμενο.

### Α.2.1 Η Διαδικασία της Επεξεργασίας Φυσικής Γλώσσας

Η κλασική διαδικασία επεξεργασίας φυσικής γλώσσας έχει ως εξής:

[2]Στη στατιστική, ένας νόμος ισχύος (power-law) είναι μια συναρτησιακή σχέση μεταξύ δύο ποσοτήτων, όπου μια σχετική μεταβολή σε μία ποσότητα έχει ως αποτέλεσμα μια σχετική μεταβολή στην άλλη ποσότητα ανάλογη με τη μεταβολή ανυψωμένη σε σταθερό εκθέτη: η μία ποσότητα μεταβάλλεται ως δύναμη της άλλης. Η μεταβολή είναι ανεξάρτητη από το αρχικό μέγεθος των ποσοτήτων αυτών [158].

**1. Κανονικοποίηση Κειμένου (Text Normalization)**

Το ακατέργαστο κείμενο διαδικτύου ή το κείμενο OCR[3] περιέχει ετερογενείς κωδικοποιήσεις, αόρατους χαρακτήρες ελέγχου και οπτικά πανομοιότυπα αλλά κανονικά διακριτά σημεία κώδικα. Η κανονικοποίηση Unicode (Unicode canonicalisation) μετατρέπει κάθε συμβολοσειρά σε μια ενιαία κανονική μορφή (συνήθως NFC[4]) έτσι ώστε, για παράδειγμα, το «é» να κωδικοποιείται πάντα ως U+00E9 αντί της αποσυντεθειμένης ακολουθίας U+0065 U+0301 [161]. Η μετατροπή σε πεζά, η αφαίρεση των σημείων στίξης και η απόκρυψη των ψηφίων μειώνουν περαιτέρω την αραιότητα για τα μοντέλα που βασίζονται σε tokens.

**2. Τμηματοποίηση (Tokenisation)**

Η διαίρεση του κειμένου σε κενά διαστήματα αποτυγχάνει για γλώσσες χωρίς ρητά όρια λέξεων (κινέζικα, ιαπωνικά, ταϊλανδικά) και για συγκολλητικές γλώσσες όπου οι μορφολογικές προσθέσεις συνενώνονται χωρίς κενά (τουρκικά, φινλανδικά). Επομένως, οι στατιστικοί ή λεξικογραφικοί διαχωριστές (segmenters) προβλέπουν τα σημεία διακοπής των tokens. Τα σύγχρονα συστήματα συχνά μαθαίνουν υπολέξεις με κωδικοποίηση Byte-Pair Encoding (BPE) ή μοντέλα γλώσσας unigram, επιτρέποντας τη δημιουργία ανοιχτού λεξιλογίου [162].

**3. Μορφολογική ανάλυση**

Πολλές γλώσσες κωδικοποιούν γραμματικές πληροφορίες (περίπτωση, αριθμός, χρόνος, γένος) ως επιθήματα (affixes). Οι *stemmers* εφαρμόζουν ευρετικές μεθόδους για την αφαίρεση επιθυμάτων ώστε να ανακτήσουν μια απλοποιημένη ή βασική μορφή μιας λέξης (π.χ. ο αλγόριθμος Porter) [163]. Οι *lemmatisers* συμβουλεύονται λεξικά και μορφολογικούς κανόνες για να επιστρέψουν το λήμμα του λεξικού και τα μορφοσυντακτικά του χαρακτηριστικά — μέρος του λόγου (Part-Of-Speech ή POS), αριθμός, πρόσωπο — διευκολύνοντας τη γενίκευση μεταξύ μορφών [159].

**4. POS επισήμανση (tagging)**

Οι επισημαντές ακολουθιών (sequence labellers) αποδίδουν συντακτικές κατηγορίες (NOUN, VERB, ADJ[5]) σε κάθε token. Τα κρυφά μοντέλα Markov, τα τυχαία πεδία υπό όρους (conditional random fields) ή τα νευρωνικά BiLSTM μοντελοποιούν τις εξαρτήσεις από τα συμφραζόμενα (contextual dependencies), επιτυγχάνοντας ακρίβεια $>97\%$ σε αγγλικά ειδησεογραφικά κείμενα [164].

**5. Συντακτική ανάλυση**

Οι αναλυτές συντακτικής (constituency parsers) δημιουργούν ένθετα φρασεολογικά δέντρα (NP ή noun phrase, VP ή verb phrase, PP ή prepositional phrase), ενώ οι αναλυτές εξάρτησης (dependency parsers) συνδέουν τις λέξεις με τόξα που εξαρτώνται από την κε-

[3] Η οπτική αναγνώριση χαρακτήρων ή οπτικός αναγνώστης χαρακτήρων (Optical Character Recognition/Reader ή OCR) είναι η ηλεκτρονική ή μηχανική μετατροπή εικόνων δακτυλογραφημένου, χειρόγραφου ή τυπωμένου κειμένου σε μηχανικά κωδικοποιημένο κείμενο, είτε από σαρωμένο έγγραφο, είτε από φωτογραφία εγγράφου, είτε από φωτογραφία σκηνής (π.χ. το κείμενο σε πινακίδες και διαφημιστικές πινακίδες σε φωτογραφία τοπίου), είτε από κείμενο υπότιτλων που έχει τοποθετηθεί σε εικόνα (π.χ. από τηλεοπτική εκπομπή) [160].

[4] Στο NFC (Normalization Form Canonical Composition) οι χαρακτήρες αποσυντίθενται και στη συνέχεια ανασυντίθενται με κανονική ισοδυναμία, έτσι ώστε τα ίδια σύμβολα να μοιράζονται ένα σημείο κώδικα.

[5] *Ουσιαστικό*, *ρήμα* και *επίθετο*, αντίστοιχα

φαλή (head-dependent arcs), π.χ. *eat* → *apples*. Τα γραφήματα εξάρτησης είναι πιο απλά και συσχετίζονται με λειτουργικές σχέσεις που είναι χρήσιμες για τη μετάφραση και την εξαγωγή σχέσεων [165].

**6. Σημασιολογική ερμηνεία**

Η αναγνώριση ονοματικών οντοτήτων (Named-Entity Recognition ή NER) ανιχνεύει και ταξινομεί επιφανειακές εκτάσεις που αντιπροσωπεύουν οντότητες του πραγματικού κόσμου (πρόσωπα, οργανισμοί, τοποθεσίες) [166]. Η σημασιολογική επισήμανση ρόλων προσδιορίζει δομές κατηγόρων – επιχειρημάτων (δράστης, υποκείμενο, μέσο) που απαντούν στο ερώτημα «ποιος έκανε τι σε ποιον» [167]. Η επίλυση συν-αναφορών (coreference resolution) συνδέει αναφορές που αναφέρονται στην ίδια οντότητα του λόγου, μια προϋπόθεση για τη συνεκτική περίληψη και την απάντηση σε ερωτήσεις [168].

Το modular pipeline εξωτερικεύει κάθε γλωσσικό επίπεδο, αλλά η σύγχρονη NLP συχνά συρρικνώνει αυτά τα βήματα σε συνεχείς διανυσματικούς χώρους που κωδικοποιούν έμμεσα ορθογραφικά, μορφολογικά, συντακτικά και σημασιολογικά χαρακτηριστικά. Οι distributional αναπαραστάσεις παρέχουν αυτό το ενοποιητικό υπόστρωμα: ο ίδιος πίνακας ενσωμάτωσης που ομαλοποιεί το surface variation μπορεί να τροφοδοτήσει POS taggers, parsers και entity recognisers χωρίς χειροκίνητο σχεδιασμό χαρακτηριστικών. Η επόμενη υποενότητα παρακολουθεί τη μετάβαση από διακριτά σύμβολα σε πυκνές ενσωματώσεις, υπογραμμίζοντας την κεντρική τους θέση στα μεγάλα γλωσσικά μοντέλα.

### Α.2.2 Σημασιολογία Κατανομής (Distributional Semantics)

Η υπόθεση της κατανομής «Θα γνωρίσεις μια λέξη από την παρέα της» (*You shall know a word by the company it keeps*) [169] οδήγησε στην ανάπτυξη πυκνών ενσωματώσεων λέξεων (dense word embeddings). Τα μοντέλα Skip-gram και CBOW στο Word2Vec βελτιστοποιούν έναν predictive στόχο που προκαλεί συνεχείς διανύσματα που καταγράφουν συντακτικές και σημασιολογικές κανονικότητες [170]. Το GloVe παραγοντοποιεί τις συνολικές co-occurence statistics, επιτυγχάνοντας παρόμοια αποτελέσματα με μια matrix-factorisation view [171].

Τα στατικά εμφυτεύματα (static embeddings) αναθέτουν ένα διάνυσμα ανά τύπο λέξης, αγνοώντας την πολυσημία. Μοντέλα υπολέξεων όπως το Byte-Pair Encoding (BPE) [162] ή το SentencePiece unigram segmentation αντιμετωπίζουν τα προβλήματα εκτός λεξιλογίου αποσυνθέτοντας τις λέξεις σε υπομονάδες που έχουν μάθει στατιστικά. Το fastText υπολογίζει επιπλέον τον μέσο όρο των $n$-grams χαρακτήρων για να συνθέσει νέες μορφές [172].

Τα contextual representations ξεπερνούν τον περιορισμό της πολυσημίας, ρυθμίζοντας τις ενσωματώσεις με βάση τα γειτονικά tokens. Το ELMo στοιβάζει αμφίδρομα LSTM και παράγει διανύσματα ειδικά για κάθε token σε κάθε επίπεδο [173]. Τα masked-language μοντέλα βασισμένα σε transformers (π.χ. BERT) βελτιώνουν την ιδέα με βαθιά self-attention, επιτυγχάνοντας μεταφορά σε κάθε σημαντικό NLP benchmark [34]. Τέτοια μοντέλα αντικαθιστούν τα διακριτά στάδια του pipeline με ενιαία εκμάθηση αναπαράστασης (representation learning), καθιστώντας την μηχανική χαρακτηριστικών (feature engineering) σε μεγάλο βαθμό παρωχημένη.

Μια ενιαία προοπτική ευθυγραμμίζεται με τις ανάγκες των εικονικών βοηθών που βασίζονται σε LLM: οι contextualised representations και το self-attention παρέχουν τη βάση για

τη δημιουργία διαλόγων, την αναγνώριση προθέσεων και τη μεταφορά σε πολλές γλώσσες.

## A.3 Αρχιτεκτονική Transformer

Ο transformer εγκαταλείπει την επαναληψιμότητα και τη συνέλιξη υπέρ της αυτοπροσοχής (self-attention), ενός content-based μηχανισμού που μετρά την επίδραση των ζευγών token σε μια ενιαία πράξη πίνακα (matrix operation) [2]. Ο παράλληλος υπολογισμός σε όλες τις θέσεις της ακολουθίας αποδίδει κορυφαία ποιότητα, ενώ ταυτόχρονα προκαλεί κορεσμό (δηλαδή πλήρης αξιοποίηση) της αριθμητικής απόδοσης του σύγχρονου υλικού GPU/TPU.

### A.3.1 Προσοχή με κλικακώμενο Εσωτερικό Γινόμενο (Scaled Dot-Product Attention)

Έχοντας αποδείξει γιατί η προσοχή υπερισχύει της επαναληψιμότητας, θα τυποποιήσουμε τώρα μαθηματικά τον μηχανισμό. Δεδομένης μιας ακολουθίας μήκους $n$ κωδικοποιημένης ως ερωτήσεων (queries ή Q), κλειδιών (keys ή K) και τιμών (values ή V), $Q, K, V \in \mathbb{R}^{n \times d_k}$, η αυτοπροσοχή επιστρέφει ένα σταθμισμένο άθροισμα τιμών:

$$\text{Attention}(Q, K, V) \;=\; \text{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right) V.$$

Ο παρονομαστής $\sqrt{d_k}$ κλιμακώνει τα logits για να διατηρήσει το μέγεθος της κλίσης (gradient magnitude) για μεγάλες διαστάσεις. Ο όρος `softmax` παράγει έναν στοχαστικό πίνακα $n \times n$ του οποίου η καταχώριση $i,j$ ποσοτικοποιεί πόσο έντονα το token $i$ ανταποκρίνεται στο token $j$. Οι ερωτήσεις, τα κλειδιά και οι τιμές προκύπτουν από την ίδια είσοδο μέσω μαθημένων γραμμικών προβολών[6].

#### Multi-head formulation

Αντί για έναν μεγάλο χάρτη προσοχής, το μοντέλο χρησιμοποιεί $h$ παράλληλες «κεφαλές» (heads), καθεμία με τους δικούς της πίνακες προβολής. Οι έξοδοι συνδέονται σειριακά και αναμιγνύονται γραμμικά:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\Big(H_1, \ldots, H_h\Big) W^O, \quad H_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right).$$

Εδώ $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ είναι μια προβολή εξόδου που αναμιγνύει τις συνενωμένες εξόδους των κεφαλών πίσω στη διάσταση του μοντέλου. Οι multi-head σχεδιασμοί καταγράφουν ετερογενείς σχέσεις, π.χ. συντακτική εξάρτηση και long-range coreference, σε διαφορετικούς υποχώρους [174].

[6] $Q$, $K$ και $V$ είναι ιστορικά ονόματα από την ανάκτηση πληροφοριών (information retrieval): ένα query αναζητά ένα σύνολο ζευγών key-value.

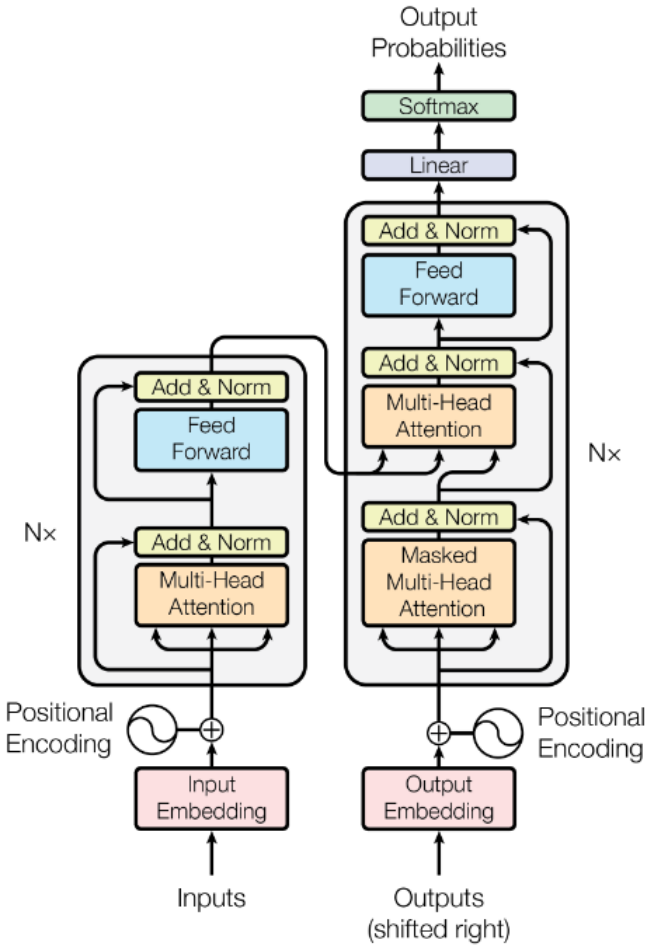
### A.3.2 Positional Encoding

Η αυτοπροσοχή είναι content-based και επομένως είναι ανεξάρτητη από τη σειρά των λέξεων. Τα σήματα θέσης (positional signals) αποκαθιστούν τη δομή της ακολουθίας. Επειδή η προσοχή είναι αμετάβλητη ως προς την μετάθεση (permutation), οι πληροφορίες θέσης εισάγονται πρόσθετα:

$$\mathrm{PE}_{\sin}(p, 2i) = \sin(p/10000^{2i/d}), \quad \mathrm{PE}_{\cos}(p, 2i+1) = \cos(p/10000^{2i/d}),$$

όπου $p$ είναι ο δείκτης της θέσης της ακολουθίας και $i$ είναι ο δείκτης της διάστασης του διανύσματος [2]. Οι ημιτονοειδείς συναρτήσεις επιτρέπουν στο μοντέλο να εξάγει συμπεράσματα για μεγαλύτερα πλαίσια, επειδή οι σχετικές αποστάσεις αντιστοιχούν σε μετατοπίσεις φάσης. Οι μαθημένες ενσωματωμένες θέσεις (positional embeddings) είναι μια εναλλακτική λύση που ανταλλάσσει την εξαγωγή συμπερασμάτων (extrapolation) με την βελτιστοποίηση για συγκεκριμένες εργασίες (task-specific optimisation) [175].

### A.3.3 Encoder-decoder stack




Σχήμα Α'.1: *Encoder-decoder transformer. Σχήμα 1 από Vaswani και συν. [2].*

Κάθε επίπεδο κωδικοποιητή (encoder) εφαρμόζει (i) αυτοπροσοχή πολλαπλών κεφαλών (multi-head self attention), (ii) ένα δίκτυο προώθησης θέσης (position-wise feed-forward

network) με δύο γραμμικά επίπεδα ενεργοποίησης ReLU/GeLU[7], και τα δύο περιτυλιγμένα από υπολειπόμενες συνδέσεις (residual connections) και κανονικοποίηση επιπέδων (layer normalization ή LN).

Το επίπεδο αποκωδικοποιητή (decoder) προσθέτει (i) κρυφή (masked) αυτοπροσοχή, εμποδίζοντας τα tokens να προσέχουν μελλοντικές θέσεις, και (ii) διασταυρούμενη προσοχή (cross-attention) της οποίας οι ερωτήσεις προέρχονται από τον αποκωδικοποιητή και τα κλειδιά/τιμές από τον κωδικοποιητή, επιτρέποντας την υπό όρους παραγωγή (μετάφραση, περίληψη).

**Πολυπλοκότητα**

Η αυτοπροσοχή κλιμακώνεται ως $O(n^2 d)$ στη μνήμη και τον υπολογισμό, ενώ τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks ή RNNs) κλιμακώνονται γραμμικά σε $n$ αλλά μόνο διαδοχικά. Για πρακτικά μήκη πλαισίου ($< 4096$) και παράλληλο υλικό, οι transformers εκπαιδεύονται γρηγορότερα και επιτυγχάνουν υψηλότερη ποιότητα. Οι πυρήνες αραιότητας (sparsity kernels) ή οι παραλλαγές γραμμικής προσοχής (linear-attention variants) μετριάζουν τον τετραγωνικό όρο για πολύ μακρές ακολουθίες [177].

### A.3.4 Παραλλαγές και ερμηνείες

Τα encoder-only μοντέλα (BERT) αποκρύπτουν (mask) τα tokens και μαθαίνουν αμφίδρομο πλαίσιο (bidirectional context) για την κατανόηση των εργασιών [34]. Τα decoder-only μοντέλα (GPT, Gemini) αυτοπαλινδρομούν (autoregress) τα προηγούμενα tokens για ανοιχτή παραγωγή (open-ended generation) [36]. Παρά την αρχιτεκτονική περικοπή (architectural pruning), όλα μοιράζονται την ίδια βασική δομή προσοχής, υπολειπόμενες διαδρομές (residual pathways) και πρόγραμμα προ-εκπαίδευσης μεγάλης κλίμακας. Οι transformers παρέχουν έτσι το υπολογιστικό υπόστρωμα πάνω στο οποίο χτίζονται τα μεγάλα γλωσσικά μοντέλα.

[7] Το ReLU (Rectified Linear Unit) επιστρέφει $\max(0, x)$. Το GeLU (Gaussian Error Linear Unit) σταθμίζει τις εισόδους με την κατανομή πιθανότητας Gauss, $x\,\Phi(x)$, βελτιώνοντας την βελτιστοποίηση σε deep transformers [176].

## Παράρτημα Β΄

# Δοκιμή τοπικής χρήσης LLaMA

Αυτό το παράρτημα περιλαμβάνει τον πλήρη κώδικα Python που χρησιμοποιήθηκε κατά την αρχική προσπάθεια τοπικής ανάπτυξης και αλληλεπίδρασης με ένα chatbot βασισμένο στο LLaMA. Η δοκιμαστική εγκατάσταση ήταν μέρος του πειράματος που περιγράφεται στην Ενότητα 4.1 Πειραματική εφαρμογή LLaMA και υλοποιήθηκε χρησιμοποιώντας το περιβάλλον εκτέλεσης *Ollama* και τη διεπαφή *LangChain*. Ο κώδικας παρουσιάζει τη σύνθεση ερωτήσεων (prompts), την κλήση μοντέλων και τη διαχείριση του περιβάλλοντος σε έναν απλό βρόχο διαλόγου γραμμής εντολών.

```
import os
from langchain_ollama import OllamaLLM
from langchain_core.prompts import ChatPromptTemplate

# Set the path to the Ollama executable
os.environ["OLLAMA_PATH"] = "C:\\Users\\juant\\AppData\\Local\\Programs\\
    Ollama\\ollama.exe"

template = """
Answer the question below.

Here is the conversation history: {context}

Question: {question}

Answer:
"""

# Initialize the model
model = OllamaLLM(model="llama3:2.1b")
prompt = ChatPromptTemplate.from_template(template)
chain = prompt | model

def handle_converation():
    context = ""
    print("Welcome to the AI Chatbot! Type 'exit' to quit.")
    while True:
        user_input = input("You: ")
        if user_input.lower() == "exit":
            break
```

```
        result = chain.invoke({"context": context, "question": user_input})
        print(f"Bot: {result}")
        context += f"\nUser: {user_input}\nAI: {result}"

if __name__ == "__main__":
    handle_converation()
```

## Παράρτημα Γ'

# Python SDK για Vertex AI Tuning

Αυτό το παράρτημα περιέχει τα αποσπάσματα κώδικα Python που χρησιμοποιούνται για την αλληλεπίδραση με την πλατφόρμα Vertex AI για την εκκίνηση και τη διαχείριση εργασιών εποπτευόμενου fine-tuning. Τα κομμάτια κώδικα που εμφανίζονται εδώ αντιστοιχούν στη διαδικασία που περιγράφεται στην Υποενότητα 4.3.1 Εκτέλεση Fine-Tuning Εργασιών και δείχνουν τη χρήση του Vertex AI SDK για τη δημιουργία, την παρακολούθηση και τον έλεγχο εργασιών.

**Δημιουργία supervised tuning job**

```python
import time

import vertexai
from vertexai.tuning import sft

# TODO(developer): Update and un-comment below line
# PROJECT_ID = "your-project-id"
vertexai.init(project=PROJECT_ID, location="us-central1")

# Initialize Vertex AI with your service account for BYOSA (Bring Your Own
    Service Account).
# Uncomment the following and replace "your-service-account"
# vertexai.init(service_account="your-service-account")

# Initialize Vertex AI with your CMEK (Customer-Managed Encryption Key).
# Un-comment the following line and replace "your-kms-key"
# vertexai.init(encryption_spec_key_name="your-kms-key")

sft_tuning_job = sft.train(
    source_model="gemini-2.0-flash-001",
    # 1.5 and 2.0 models use the same JSONL format
    train_dataset="gs://cloud-samples-data/ai-platform/generative_ai/gemini
        -1_5/text/sft_train_data.jsonl",
    # The following parameters are optional
    validation_dataset="gs://cloud-samples-data/ai-platform/generative_ai/
        gemini-1_5/text/sft_validation_data.jsonl",
    tuned_model_display_name="tuned_gemini_2_0_flash",
    # Advanced use only below. It is recommended to use auto-selection and
        leave them unset
    # epochs=4,
```

```
    # adapter_size=4,
    # learning_rate_multiplier=1.0,
)

# Polling for job completion
while not sft_tuning_job.has_ended:
    time.sleep(60)
    sft_tuning_job.refresh()

print(sft_tuning_job.tuned_model_name)
print(sft_tuning_job.tuned_model_endpoint_name)
print(sft_tuning_job.experiment)
# Example response:
# projects/123456789012/locations/us-central1/models/12345678901
# projects/123456789012/locations/us-central1/endpoints/123456789012345
# <google.cloud.aiplatform.metadata.experiment_resources.Experiment object
    at 0x7b5b4ae07af0>
```

**Λίστα από tuning jobs**

```
import vertexai
from vertexai.tuning import sft

# TODO(developer): Update and un-comment below line
# PROJECT_ID = "your-project-id"
vertexai.init(project=PROJECT_ID, location="us-central1")

responses = sft.SupervisedTuningJob.list()

for response in responses:
    print(response)
# Example response:
# <vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0
    x7c85287b2680>
# resource name: projects/12345678/locations/us-central1/tuningJobs
    /123456789012345
```

**Λεπτομέρειες ενός tuning job**

```
import vertexai
from vertexai.tuning import sft

# TODO(developer): Update and un-comment below lines
# PROJECT_ID = "your-project-id"
# LOCATION = "us-central1"
vertexai.init(project=PROJECT_ID, location=LOCATION)

tuning_job_id = "4982013113894174720"
response = sft.SupervisedTuningJob(
    f"projects/{PROJECT_ID}/locations/{LOCATION}/tuningJobs/{tuning_job_id}
        "
)
```

```
print(response)
# Example response:
# <vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0
    x7cc4bb20baf0>
# resource name: projects/1234567890/locations/us-central1/tuningJobs
    /4982013113894174720
```

## Παράρτημα Δ΄

# Αντιστοίχιση λέξεων-κλειδιών με προτροπές

Αυτό το παράρτημα περιέχει τον πλήρη κατάλογο των αναγνωριστικών λέξεων-κλειδιών (keywords) που χρησιμοποιούνται στη διαδικασία επαύξησης προτροπών με αναγνώριση περιβάλλοντος που περιγράφεται στην Υποενότητα 5.2.2 Τροποποίηση του συνόλου δεδομένων με λέξεις-κλειδιά. Κάθε λέξη-κλειδί αντιστοιχεί σε μια συγκεκριμένη καρτέλα της διεπαφής της πλατφόρμας *ENERGATE*. Κατά τη διάρκεια του fine-tuning, αυτές οι λέξεις-κλειδιά προστέθηκαν στις προτροπές για να προσομοιώσουν την ευαισθησία του βοηθού στην τοποθεσία του χρήστη.

Για κάθε λέξη-κλειδί, παρέχεται ένα παράδειγμα προτροπής για να απεικονιστεί ο τρόπος με τον οποίο ενσωματώθηκε στα δεδομένα εκπαίδευσης. Παρόλα αυτά, μία λέξη-κλειδί μπορεί να έχει χρησιμοποιηθεί σε διαφορετικές ερωτήσεις, ενώ μία ερώτηση μπορεί να αντιστοιχεί σε πολλές (διαφορετικές) λέξεις-κλειδιά.

*Αντιστοίχιση λέξεων-κλειδιών και προτροπών στην τελική ενίσχυση του μοντέλου*

| **Λέξη-Κλειδί** | **Παράδειγμα Ερώτησης** |
|---|---|
| `homepage` | What is the ENERGATE platform? |
| `private_building_details` | What details does a Building Owner need to provide in the tab 'Builiding Details' when adding a new building? |
| `private_building_typology` | What details does a Building Owner need to provide in the tab 'Builiding Typology' when adding a new building? |
| `private_building_technical` | What details does a Building Owner need to provide in the tab 'Technical Information' when adding a new building? |
| `private_building_energyefficiency` | What details does a Building Owner need to provide in the tab 'Energy Efficiency Status' when adding a new building? |
| `private_building_financing` | What details does a Building Owner need to provide in the tab 'Project Financing' when adding a new building? |

| **Λέξη-Κλειδί** | **Παράδειγμα Ερώτησης** |
|---|---|
| private_building_preferredmeasures | What details does a Building Owner need to provide in the tab 'Preferred Measures' when adding a new building? |
| currentprojects | How does a Building Owner update or delete a building? |
| private_building_initialoffer | What does the initial offer presented to a Building Owner include? |
| addbuilding_finaloffer | How many steps does the final offer presented to a Building Owner include? |
| detailedoffer_finaloffer | How many steps are involved in the final offer presented to a Building Owner? |
| addbuilding_finaloffer1 | What does the step one of final offer presented to a Buiding Owner include? |
| addbuilding_finaloffer2 | Can you describe what step one of the final offer to a Building Owner entails? |
| addbuilding_finaloffer3 | What does the step two of final offer presented to a Buiding Owner include? |
| matches | What does the dashboard of the Implementors include? |
| detailedoffer_calculations | What economic indicators does the platform calculate? |
| addbuilding | How does an Implementor add a new building? |
| addbuilding_details | What details does an Implementor need to provide in the tab 'Building details' when adding a new building? |
| addbuilding_typology | What details does an Implementor need to provide in the tab 'Building typology' when adding a new building? |
| addbuilding_technical | What details does an Implementor need to provide in the tab 'Technical information' when adding a new building? |
| addbuilding_energyefficiency | What details does an Implementor need to provide in the tab 'Energy Efficiency Status' when adding a new building? |
| addbuilding_financing | What details does an Implementor need to provide in the tab 'Project financing' when adding a new building? |
| createclusters | What does the tab 'Create Clusters' do? |
| account | Where can an Implementor see their account information? |
| building | What steps should a Public Body follow to add a new building? |

| **Λέξη-Κλειδί** | **Παράδειγμα Ερώτησης** |
|---|---|
| `public_building_details` | What details does a Public Body need to provide in the tab 'Building Details' when adding a new building? |
| `public_building_typology` | What details does a Public Body need to provide in the tab 'Building typology' when adding a new building? |
| `public_building_technical` | What details does a Public Body need to provide in the tab 'Technical information' when adding a new building? |
| `public_building_energyefficiency` | What details does a Public Body need to provide in the tab 'Energy Efficiency Status' when adding a new building? |
| `public_building_financing` | What details does a Public Body need to provide in the tab 'Project financing' when adding a new building? |

# Bibliography


[1] R. Bommasani *et al.* *On the Opportunities and Risks of Foundation Models*. arXiv preprint arXiv:2108.07258, 2021.

[2] A. Vaswani *et al.* *Attention Is All You Need*. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2017.

[3] H. Ritchie, M. Roser. *CO2 and Greenhouse Gas Emissions*, 2024. Accessed: 2025-06-27.

[4] European Commission. *Directive (EU) 2023/1791 on Energy Efficiency*. *Official Journal of the European Union*, 2023.

[5] European Commission. *Recast of the Energy Performance of Buildings Directive*. COM(2024) XXX final, 2024.

[6] J. Doe *et al.* *Cost-Benefit Analysis of Sustainable Upgrades in Existing Buildings*. *Energy and Buildings*, 2024.

[7] N. Park *et al.* *Energy Savings and Economics of Retrofitting Single-Family Buildings*. Τεχνική Αναφορά με αριθμό, Lawrence Berkeley National Laboratory, 2025.

[8] ENERGATE Consortium. *ENERGATE: Smart Marketplace for Energy Efficiency Investments*, 2024. Accessed: 2025-06-27.

[9] Institutefor European Energy, Climate Policy (IEECP). *ENERGATE - Energy Efficiency Aggregation Platform for Sustainable Investments*, 2025. Accessed: 2025-03-15.

[10] A. Smith, P. García. *Digital Tools for Stakeholder Participation in Urban Development*. *Smart Cities*, 2023.

[11] McKinsey & Company. *What are Industry 4.0, the Fourth Industrial Revolution, and 4IR?*, 2023. Accessed: 2025-04-30.

[12] D. Galanakis, G. Kambourakis, E. Chatzoglou. *Virtual Assistants in Industry 4.0: A Systematic Literature Review*. *Electronics*, 12(19):4096, 2023.

[13] *LLaMA's Hugging Face Repository*. https://huggingface.co/docs/transformers/en/model_doc/llama, 2025. Accessed: 2025-03-15.

[14] *BLOOM-176B's Hugging Face Repository*. https://huggingface.co/docs/transformers/en/model_doc/bloom, 2025. Accessed: 2025-03-15.

[15] *OpenAI's GPT2 Hugging Face Repository*. https://huggingface.co/openai-community/gpt2, 2025. Accessed: 2025-03-15.

[16] *Gemini Google DeepMind*. https://deepmind.google/technologies/gemini/, 2025. Accessed: 2025-03-15.

[17] S. Kothapalli, P. Kalyani, P. Chandrika. *Chatbots and Virtual Assistants in HRM: Exploring Their Role in Employee Engagement and Support*. https://www.researchgate.net/publication/383360149, 2024. Accessed: 2025-04-13.

[18] A. P. Chaves, M. A. Gerosa. *Customer service chatbots: Anthropomorphism and adoption*.

*Journal of Business Research*, 123:100–112, 2021.

[19] S. Kim. *Ethical Challenges in the Development of Virtual Assistants Powered by LLMs*. *Electronics*, 2024.

[20] M. Adam, B. Wessel, M. Stein. *Service chatbots: A systematic review*. *Expert Systems with Applications*, 165:113918, 2021.

[21] Wikipedia contributors. *F-score - Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/F-score, 2024. Accessed: 2025-04-27.

[22] Wikipedia contributors. *BLEU - Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/BLEU, 2024. Accessed: 2025-04-27.

[23] A. K. Sharma, R. Prasad. *AI Chatbots: Transforming the Digital World*. *Proc. Int. Conf. on Innovative Computing and Communications*A. Kumar, R. P. Maheshwari, H. S. Behera, P. K. Pattnaik, επιμελητές, σελίδες 407–417. Springer, 2020.

[24] S. J. Russell, P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4η έκδοση, 2021.

[25] J. McCarthy, M. Minsky, C. Shannon, N. Rochester. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. https://en.wikipedia.org/wiki/Dartmouth_workshop, 1956. Accessed: 2025-06-27.

[26] A. L. Samuel. *Some Studies in Machine Learning Using the Game of Checkers*. *IBM Journal of Research and Development*, 1959.

[27] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[28] I. Goodfellow, Y. Bengio, A. Courville. *Deep Learning*. MIT Press, 2016.

[29] Wikipedia contributors. *Symbolic artificial intelligence*. https://en.wikipedia.org/wiki/Symbolic_artificial_intelligence, 2025. Accessed: 2025-06-27.

[30] M. I. Jordan, T. M. Mitchell. *Machine Learning: Trends, Perspectives, and Prospects*. *Science*, 349(6245):255–260, 2015.

[31] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[32] J. MacQueen. *Some Methods for Classification and Analysis of Multivariate Observations*. *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, Vol. 1: Statistics*, σελίδες 281–297, Berkeley, CA, USA, 1967. University of California Press.

[33] *Downstream Task Definition*. https://www.baeldung.com/cs/downstream-tasks, 2025. Accessed: 2025-03-15.

[34] J. Devlin *et al.* *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *arXiv preprint arXiv:1810.04805*, 2018.

[35] R. S. Sutton, A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2η έκδοση, 2018.

[36] T. B. Brown *et al.* *Language Models are Few-Shot Learners*. arXiv preprint arXiv:2005.14165, 2020.

[37] J. Kaplan *et al.* *Scaling Laws for Neural Language Models*. *arXiv preprint arXiv:2001.08361*, 2020.

[38] J. Hoffmann *et al.* *Training Compute-Optimal Large Language Models*. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[39] S. Pichai, D. Hassabis. *Introducing Gemini: Our Largest and Most Capable AI Model*. https://blog.google/technology/ai/google-gemini-ai, 2025. Accessed: 2025-03-15.

[40] H. W. Chung *et al.* *Scaling Instruction-Finetuned Language Models*. arXiv preprint arXiv:2210.11416, 2022.

[41] L. Lee *et al.* *Deduplicating Training Data Makes Language Models Better*. arXiv preprint arXiv:2107.06499, 2021.

[42] C. Raffel *et al.* *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR Workshop and Conference Proceedings*, 2020.

[43] J. Wei *et al.* *Finetuned Language Models Are Zero-Shot Learners. Proc. Int. Conf. on Learning Representations (ICLR)*, 2022.

[44] L. Ouyang *et al.* *Training Language Models to Follow Instructions with Human Feedback. Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[45] P. Lewis *et al.* *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[46] P. Micikevicius *et al.* *Mixed Precision Training. Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.

[47] S. Rajbhandari *et al.* *ZeRO: Memory Optimization Toward Training A Trillion Parameter Models. Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis (SC)*, 2020.

[48] E. J. Hu *et al.* *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685, 2021.

[49] T. Dettmers *et al.* *8-Bit Optimizers via Blockwise Quantization*. arXiv preprint arXiv:2208.07339, 2022.

[50] E. Frantar *et al.* *GPTQ: Accurate Post-Training Quantization for Generative Transformers*. arXiv preprint arXiv:2210.17323, 2023.

[51] S. Chen *et al.* *TensorRT-LLM: High-Performance Large Language Model Inference on GPUs*. arXiv preprint arXiv:2309.03049, 2023.

[52] K. Papineni *et al.* *BLEU: A Method for Automatic Evaluation of Machine Translation. Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[53] C. Y. Lin. *ROUGE: A Package for Automatic Evaluation of Summaries. Proc. ACL Workshop on Text Summarization Branches Out*, 2004.

[54] T. Zhang *et al.* *BERTScore: Evaluating Text Generation with BERT. Proc. Int. Conf. on Learning Representations (ICLR)*, 2020.

[55] D. Hendrycks *et al.* *Measuring Massive Multitask Language Understanding*. arXiv preprint arXiv:2009.03300, 2021.

[56] A. Srivastava *et al.* *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*. arXiv preprint arXiv:2206.04615, 2022.

[57] T. Bolukbasi *et al.* *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016.

[58] J. Buolamwini, T. Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proc. Conf. on Fairness, Accountability and Transparency (FAT*)*, 2018.

[59] N. Mehrabi *et al.* *A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys*, 2021.

[60] N. Carlini *et al.* *Extracting Training Data from Large Language Models*. arXiv preprint

arXiv:2012.07805, 2021.

[61] L. Lyu *et al. Differentially Private Language Modelling. Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[62] E. Strubell *et al. Energy and Policy Considerations for Deep Learning in NLP. Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[63] *Annex III: High-Risk AI Systems Referred to in Article 6(2).* https://www.euaiact.com/annex/3, 2024. Accessed: 2025-06-27.

[64] D. Ganguli *et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.* arXiv preprint arXiv:2210.09261, 2022.

[65] R. Gao *et al. Self-Rewarding Language Models.* arXiv preprint arXiv:2309.00663, 2023.

[66] M. Galanakis, S. Mitropoulos, A. Vafeiadis. *Industry 4.0 Virtual Assistants: Enhancing Human-Machine Interaction. Journal of Intelligent Manufacturing Systems*, 35(4):225–239, 2023.

[67] A. P. Chaves, M. A. Gerosa. *Empirical Studies on Human-Chatbot Interaction: A Systematic Literature Review. Information Processing & Management*, 54(5):1–19, 2019.

[68] Google Cloud. *Vertex AI Agent Builder Documentation.* https://cloud.google.com/vertex-ai/docs/agent-builder, 2024. Accessed: 2025-06-27.

[69] *Falcon's Hugging Face Repository.* https://huggingface.co/docs/transformers/en/model_doc/falcon, 2025. Accessed: 2025-03-15.

[70] *GPT-J's Hugging Face Repository.* https://huggingface.co/docs/transformers/en/model_doc/gptj, 2025. Accessed: 2025-03-15.

[71] *GPT-NeoX's Hugging Face Repository.* https://huggingface.co/docs/transformers/en/model_doc/gpt_neox, 2025. Accessed: 2025-03-15.

[72] *Vicuna-7B's Hugging Face Repository.* https://huggingface.co/lmsys/vicuna-7b-v1.5, 2025. Accessed: 2025-03-15.

[73] *XGen-7B/8K's Hugging Face Repository.* https://huggingface.co/Salesforce/xgen-7b-8k-base, 2025. Accessed: 2025-03-15.

[74] PyTorch Team. *PyTorch - An Open Source Machine Learning Framework.* https://pytorch.org/, 2024. Accessed: 2025-03-15.

[75] Hugging Face. *Transformers Documentation.* https://huggingface.co/docs/transformers/en/index, 2024. Accessed: 2025-03-15.

[76] *Official Hugging Face Website.* https://huggingface.co/, 2025. Accessed: 2025-03-15.

[77] *BLOOM's Wikipedia Page.* https://en.wikipedia.org/wiki/BLOOM_(language_model), 2025. Accessed: 2025-03-15.

[78] BigScience. *BigScience Language Open-science Open-access Multilingual (BLOOM) Language Model.* https://bigscience.huggingface.co/, 2022. International, May 2021–May 2022.

[79] *BLOOM-560M's Hugging Face Repository.* https://huggingface.co/bigscience/bloom-560m, 2025. Accessed: 2025-03-15.

[80] *BLOOM-1.1B's Hugging Face Repository.* https://huggingface.co/bigscience/bloom-1b1, 2025. Accessed: 2025-03-15.

[81] *BLOOM-1.7B's Hugging Face Repository.* https://huggingface.co/bigscience/bloom-1b7, 2025. Accessed: 2025-03-15.

[82] *BLOOM-3B's Hugging Face Repository*. https://huggingface.co/bigscience/bloom-3b, 2025. Accessed: 2025-03-15.

[83] *BLOOM-7.1B's Hugging Face Repository*. https://huggingface.co/bigscience/bloom-7b1, 2025. Accessed: 2025-03-15.

[84] Hugging Face. *Tokenizers - Hugging Face Documentation*. https://huggingface.co/docs/transformers/tokenizer_summary, 2024. Accessed: 2025-03-15.

[85] BigScience Workshop, T. Le Scao, A. Fan, C. Akiki, E. Pavlick *et al.* *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. https://arxiv.org/abs/2211.05100, 2023.

[86] *TII, Falcon's Development Team*. https://www.tii.ae/, 2025. Accessed: 2025-03-15.

[87] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru *et al.* *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*. https://arxiv.org/abs/2306.01116, 2023.

[88] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer *et al.* *Open LLM Leaderboard v2*. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024. Accessed: 2025-03-25.

[89] *Falcon-7B's Hugging Face Repository*. https://huggingface.co/tiiuae/falcon-7b, 2025. Accessed: 2025-03-15.

[90] *Falcon-40B's Hugging Face Repository*. https://huggingface.co/tiiuae/falcon-40b, 2025. Accessed: 2025-03-15.

[91] *Falcon-180B's Hugging Face Repository*. https://huggingface.co/tiiuae/falcon-180B, 2025. Accessed: 2025-03-15.

[92] E. Nijkamp, T. Xie, H. Hayashi, B. Pang *et al.* *The Falcon Series of Language Models: Towards Open Frontier Models*. https://arxiv.org/abs/2309.03450, 2023.

[93] M. Tyagi. *Gemini: Google's Multimodal Mastermind and Its Impact on AI*. https://www.linkedin.com/pulse/gemini-googles-multimodal-mastermind-its-impact-ai-manish-tyagi-du2tf/, 2025. Accessed: 2025-03-15.

[94] *Google Cloud Vertex AI Platform*. https://cloud.google.com/vertex-ai?hl=en, 2025. Accessed: 2025-03-15.

[95] *OpenAI's GPT2 Page*. https://openai.com/index/gpt-2-1-5b-release/, 2025. Accessed: 2025-03-15.

[96] *OpenAI's GPT2 Wikipedia Page*. https://en.wikipedia.org/wiki/GPT-2, 2025. Accessed: 2025-03-15.

[97] *PILE Dataset (Used on GPT-J and GPT-NeoX)*. https://pile.eleuther.ai/, 2025. Accessed: 2025-03-15.

[98] *GPT-J's GitHub Repository*. https://github.com/kingoflolz/mesh-transformer-jax, 2025. Accessed: 2025-03-15.

[99] TensorFlow Authors. *TensorFlow - An End-to-End Open Source Machine Learning Platform*. https://www.tensorflow.org/, 2024. Accessed: 2025-03-15.

[100] Flax Developers. *Flax Documentation*. https://flax.readthedocs.io/en/latest/, 2024. Accessed: 2025-03-15.

[101] A. Komatsuzaki. *GPT-J-6B: 6B JAX-Based Transformer*. https://arankomatsuzaki.wordpress.com/2021/06/04/gpt-j/, 2021. Accessed: 2025-03-15.

[102] *EleutherAI, GPT-NeoX's Development Team*. https://www.eleuther.ai/, 2025. Accessed: 2025-03-15.

[103] *CoreWeave, GPT-NeoX's Supporting Team on Training*. https://www.coreweave.com/, 2025. Accessed: 2025-03-15.

[104] *OpenAI's GPT3 Page*. https://openai.com/index/gpt-3-apps/, 2025. Accessed: 2025-03-15.

[105] *OpenAI's GPT3 Wikipedia Page*. https://en.wikipedia.org/wiki/GPT-3, 2025. Accessed: 2025-03-15.

[106] *Meta's FairSeq Blogpost*. https://ai.meta.com/research/publications/fairseq-a-fast-extensible-toolkit-for-sequence-modeling/, 2025. Accessed: 2025-03-15.

[107] *GPT-NeoX's GitHub Repository*. https://github.com/EleutherAI/gpt-neox, 2025. Accessed: 2025-03-15.

[108] *LLaMA's Wikipedia Page*. https://en.wikipedia.org/wiki/Llama_(language_model), 2025. Accessed: 2025-03-15.

[109] H. Touvron, T. Lavril, G. Izacard, X. Martinet *et al.* *LLaMA: Open and Efficient Foundation Language Models*. https://arxiv.org/abs/2302.13971, 2023. Accessed: 2025-03-15.

[110] *LLaMA2's Hugging Face Repository*. https://huggingface.co/docs/transformers/en/model_doc/llama2, 2025. Accessed: 2025-03-15.

[111] H. Touvron, L. Martin, K. Stone, P. Albert *et al.* *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/, 2023. Meta AI.

[112] *LLaMA3's Hugging Face Repository*. https://huggingface.co/docs/transformers/en/model_doc/llama3, 2025. Accessed: 2025-03-15.

[113] *Meta's Blogpost for LLaMA3 (No Research Paper Available)*. https://ai.meta.com/blog/meta-llama-3/, 2025. Accessed: 2025-03-15.

[114] *Chinchilla's Wikipedia Page*. https://en.wikipedia.org/wiki/Chinchilla_(language_model), 2025. Accessed: 2025-03-15.

[115] *PaLM's Wikipedia Page*. https://en.wikipedia.org/wiki/PaLM, 2025. Accessed: 2025-03-15.

[116] *Google Colaboratory Page*. https://colab.google/, 2025. Accessed: 2025-03-15.

[117] *LLaMA Simple Finetuner Jupyter Notebook*. https://colab.research.google.com/github/lxe/simple-llama-finetuner/blob/master/Simple_LLaMA_FineTuner.ipynb#scrollTo=3PM_DilAZD8T, 2025. Accessed: 2025-03-15.

[118] *LLaMA Amazon SageMaker Jupyter Notebook*. https://github.com/aws/amazon-sagemaker-examples/blob/main/introduction_to_amazon_algorithms/jumpstart-foundation-models/text-generation-open-llama.ipynb, 2025. Accessed: 2025-03-15.

[119] *LMSYS, Vicuna's Development Team*. https://lmsys.org/, 2025. Accessed: 2025-03-15.

[120] *ShareGPT, Vicuna's Dataset Source (Deprecated)*. https://sharegpt.com/, 2025. Accessed: 2025-03-15.

[121] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang *et al.* *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. https://arxiv.org/abs/2306.05685, 2023.

[122] Vicuna Team. *Vicuna Blogpost*. https://lmsys.org/blog/2023-03-30-vicuna/, 2025.

Accessed: 2025-03-15.

[123] *Vicuna-7B's Command Line Interface*. https://github.com/lm-sys/FastChat#vicuna-weights, 2025. Accessed: 2025-03-15.

[124] *Vicuna's GitHub Repository*. https://github.com/lm-sys/FastChat, 2025. Accessed: 2025-03-15.

[125] *Vicuna Demo Interface*. https://lmarena.ai/, 2025. Accessed: 2025-03-15.

[126] *Salesforce's Blogpost on XGen*. https://www.salesforce.com/blog/xgen/, 2025. Accessed: 2025-03-15.

[127] *NLP Wikipedia Page*. https://en.wikipedia.org/wiki/Natural_language_processing, 2025. Accessed: 2025-03-15.

[128] *NLP IBM Page*. https://www.ibm.com/think/topics/natural-language-processing, 2025. Accessed: 2025-03-15.

[129] *MPT Hugging Face Repository*. https://huggingface.co/docs/transformers/en/model_doc/mpt, 2025. Accessed: 2025-03-15.

[130] *Redpajama GitHub Repository*. https://github.com/togethercomputer/RedPajama-Data, 2025. Accessed: 2025-03-15.

[131] *OpenLLaMA GitHub Repository*. https://github.com/openlm-research/open_llama, 2025. Accessed: 2025-03-15.

[132] *XGen-7B/4K's Hugging Face Repository*. https://huggingface.co/Salesforce/xgen-7b-4k-base, 2025. Accessed: 2025-03-15.

[133] E. Nijkamp, T. Xie, H. Hayashi, B. Pang *et al. Long Sequence Modeling with XGen: A 7B LLM Trained on 8K Input Sequence Length*. https://arxiv.org/abs/2309.03450, 2023.

[134] *Likert Scale Wikipedia Page*. https://en.wikipedia.org/wiki/Likert_scale, 2025. Accessed: 2025-03-15.

[135] *Eleuther AI Language Model Evaluation Harness*. https://www.eleuther.ai/projects/large-language-model-evaluation, 2025. Accessed: 2025-03-15.

[136] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, L. Hou. *Instruction-Following Evaluation for Large Language Models*. https://arxiv.org/abs/2311.07911, 2023. arXiv:2311.07911 [cs.CL].

[137] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, J. Wei. *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them*. https://arxiv.org/abs/2210.09261, 2022. arXiv:2210.09261 [cs.CL].

[138] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt. *Measuring Mathematical Problem Solving With the MATH Dataset*. https://arxiv.org/abs/2103.03874, 2021. arXiv:2103.03874 [cs.LG].

[139] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, S. R. Bowman. *GPQA: A Graduate-Level Google-Proof QnA Benchmark*. https://arxiv.org/abs/2311.12022, 2023. arXiv:2311.12022 [cs.AI].

[140] Z. Sprague, X. Ye, K. Bostrom, S. Chaudhuri, G. Durrett. *MuSR: Testing the Limits of Chain-of-Thought with Multistep Soft Reasoning*. https://arxiv.org/abs/2310.16049, 2024. arXiv:2310.16049 [cs.CL].

[141] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang,

T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, W. Chen. *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*. https://arxiv.org/abs/2406.01574, 2024. arXiv:2406.01574 [cs.CL].

[142] *Ollama - Run Large Language Models Locally*. https://ollama.com/, 2025. Accessed: 2025-05-21.

[143] *Ollama GitHub Repository*. https://github.com/ollama/ollama, 2025. Accessed: 2025-05-21.

[144] S. Vij. *LLM Quantization: GPTQ, QAT, AWQ, GGUF, GGML, PTQ*. https://medium.com/@siddharth.vij10/llm-quantization-gptq-qat-awq-gguf-ggml-ptq-2e172cd1b3b5, 2024. Accessed: 2025-05-21.

[145] *Vertex AI Model Tuning*. https://console.cloud.google.com/vertex-ai/studio/tuning, 2025. Accessed: 2025-05-21.

[146] *Document Understanding for Gemini Tuning*. https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune_gemini/doc_tune, 2025. Accessed: 2025-05-21.

[147] jsonlines.org. *JSON Lines*, 2025. Accessed: 2025-06-04.

[148] OpenAI. *OpenAI*. https://openai.com/, 2024. Accessed: 2025-06-13.

[149] *Tune Gemini Models by Using Supervised Fine-Tuning*. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning, 2025. Accessed: 2025-05-21.

[150] *Google Cloud Storage - Browser Console*. https://console.cloud.google.com/storage/browser, 2025. Accessed: 2025-05-21.

[151] L. Bottou. *Stochastic Gradient Descent Tricks*. *Neural Networks: Tricks of the Trade*, σελίδες 421–436. Springer, 2012.

[152] A. Krogh, J. A. Hertz. *A Simple Weight Decay Can Improve Generalization*. *Advances in Neural Information Processing Systems 4*, σελίδες 950–957, 1992.

[153] L. Prechelt. *Early Stopping—But When?* *Neural Networks: Tricks of the Trade*, τόμος 1524 στο *Lecture Notes in Computer Science*, σελίδες 55–69. Springer, 1998.

[154] N. Srivastava *et al.* *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. *Journal of Machine Learning Research*, 2014.

[155] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[156] Y. LeCun, Y. Bengio, G. Hinton. *Deep Learning*. *Nature*, 521:436–444, 2015.

[157] D. E. Rumelhart, G. E. Hinton, R. J. Williams. *Learning Internal Representations by Error Propagation*. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*D. Rumelhart, J. McClelland, επιμελητές, τόμος 1, σελίδες 318–362. MIT Press, 1986.

[158] Wikipedia contributors. *Power law — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Power_law, 2025. Accessed: 2025-04-30.

[159] D. Jurafsky, J. H. Martin. *Speech and Language Processing*. Pearson, 3η έκδοση, 2023.

[160] Wikipedia contributors. *Optical character recognition — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Optical_character_recognition. Accessed: 2025-04-30.

[161] Unicode Consortium. *Unicode Standard, Version 15.1*. https://www.unicode.org/versions/Unicode15.1.0/, 2024. Accessed: 2025-06-27.

[162] R. Sennrich *et al.* *Neural Machine Translation of Rare Words with Subword Units*. *Proc.*

*Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[163] M. F. Porter. *An Algorithm for Suffix Stripping*. *Program*, 1980.

[164] K. Toutanova *et al.* *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. *Proc. NAACL*, 2003.

[165] S. Kübler, R. McDonald, J. Nivre. *Dependency Parsing*. Morgan & Claypool, 2009.

[166] E. F. Tjong Kim Sang, F. De Meulder. *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. *Proc. Conf. on Computational Natural Language Learning (CoNLL)*, 2003.

[167] D. Gildea, D. Jurafsky. *Automatic Labeling of Semantic Roles*. *Computational Linguistics*, 2002.

[168] V. Ng. *Improving Machine Learning Approaches to Coreference Resolution*. *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[169] J. R. Firth. *A Synopsis of Linguistic Theory 1930-1955*. *Studies in Linguistic Analysis*, 1957.

[170] T. Mikolov *et al.* *Efficient Estimation of Word Representations in Vector Space*. *Proc. Int. Conf. on Learning Representations (ICLR)*, 2013.

[171] J. Pennington, R. Socher, C. D. Manning. *GloVe: Global Vectors for Word Representation*. *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[172] P. Bojanowski *et al.* *Enriching Word Vectors with Subword Information*. *Transactions of the Association for Computational Linguistics*, 2017.

[173] M. Peters *et al.* *Deep Contextualized Word Representations*. *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[174] K. Clark *et al.* *What Does BERT Look at? An Analysis of BERT's Attention*. *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[175] O. Press *et al.* *ALiBi: Position Information with Linear Biases Enables Longer Contexts*. arXiv preprint arXiv:2108.12409, 2021.

[176] D. Hendrycks, K. Gimpel. *Gaussian Error Linear Units (GELUs)*. https://arxiv.org/abs/1606.08415, 2016. Accessed: 2025-06-27.

[177] Y. Tay *et al.* *Efficient Transformers: A Survey*. arXiv preprint arXiv:2009.06732, 2020.

# Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

| | |
|---|---|
| Εικονικός Βοηθός | ΕΒ |
| Ευρωπαϊκή Ένωση | ΕΕ |
| και άλλα | κ.α. |
| Μετάφραση | Μτφ. |
| Μηχανική Μάθηση | ΜΜ |
| Συμπράξεις Δημοσίου και Ιδιωτικού Τομέα | ΣΔΙΤ |
| Τεχνητή Νοημοσύνη | ΤΝ |

| | |
|---|---|
| Artificial Inteligence | AI |
| Beyond the Imitation Game Benchmark | BIG-Bench |
| Big Bench Hard | BBH |
| BigScience Large Open-science Open-access Multilingual Language Model | BLOOM |
| BiLingual Evaluation Understudy | BLEU |
| Energy Management Systems | EMS |
| Energy Perfomance Certificate | EPC |
| Energy Perfomance COmpany | ESCO |
| Energy Perfomance of Buildings Directive | EPBD |
| European Union | EU |
| Fourth Industrial Revolution | 4IR |
| Frequently Asked Questions | FAQ |
| General Data Protection Regulation | GDPR |
| Google Cloud Storage | GCS |
| Graduate-Level Google-Proof QnA | GPQA |
| Graphical User Interface | GUI |
| Heating, Ventilation and Air Conditioning | HVAC |
| Instruction-Following Evaluation | IFEval |
| Large Language Model | LLM |
| Large Language Model Meta AI | LLaMA |
| L'Instrument Financier pour L'Enviroment | LIFE |
| Low Rank Adapters | LoRA |
| Machine Learning | ML |
| Masked-Language Modelling | MLM |
| Massive Multitask Language Understanding | MMLU |
| Massive Multitask Language Understanding Professional | MMLU-Pro |

| | |
|---|---|
| Mathematics Aptitude Test of Heuristics | MATH |
| Multistep Soft Reasoning | MuSR |
| Name Entity Recognition | NER |
| Natural Language Generation | NLG |
| Natural Language Processing | NLP |
| Natural Language Understanding | NLU |
| Need For Human Interaction | NFHI |
| Out of Scope | OoS |
| Portable Document File | PDF |
| Proximal Policy Optimasation | PPO |
| Recall-Oriented Understady for Gisting Evaluation | ROUGE |
| Reinforcement Learning From Human Feedback | RLHF |
| Retrieval-Augmented Generation | RAG |
| Supervised Fine-Tuning | SFT |
| Uniform Resource Identifier | URI |
| Virtual Agent | VA |
| Zero Redundancy Optimizer | ZeRO |