Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

Data Acquisition, Exploration and Preparation for LLM Training

The Case of the Greek Language

Diploma Thesis

by

Konstantinos S. Divriotis

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

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

Data Acquisition, Exploration and Preparation for LLM Training

The Case of the Greek Language

Diploma Thesis

by

Konstantinos S. Divriotis

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2η Ιουλίου 2025.

.....	.....	.....
Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.	Μιχάλης Βαζιργιάννης Καθηγητής Ecole Polytechnique, Επισκέπτης Καθηγητής Ε.Μ.Π.	Αθανάσιος Βουλόδημος Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

ΚΩΝΣΤΑΝΤΙΝΟΣ Σ. ΔΙΒΡΙΩΤΗΣ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) έχουν αναδειχθεί ως ισχυρά εργαλεία για την Επεξεργασία Φυσικής Γλώσσας, καθοδηγούμενα από τη διαρκώς αυξανόμενη κλίμακα των μοντέλων και των συνόλων δεδομένων εκπαίδευσης. Αν και τέτοιοι πόροι είναι διαθέσιμοι για γλώσσες υψηλών πόρων, οι γλώσσες χαμηλών πόρων όπως η Ελληνική παραμένουν σημαντικά υποεκπροσωπούμενες στη σύγχρονη έρευνα και ανάπτυξη LLMs.

Στην παρούσα εργασία, επιδιώκουμε να καλύψουμε αυτό το κενό κατασκευάζοντας δύο θεμελιώδη σύνολα δεδομένων για την ανάπτυξη Ελληνικών LLMs: ένα σύνολο δεδομένων προεκπαίδευσης και ένα σύνολο για εκπαίδευση βάσει οδηγιών (instruction tuning). Για την προεκπαίδευση, συλλέξαμε και επεξεργαστήκαμε μεγάλους όγκους συνομιλιακών δεδομένων από απομαγνητοφωνήσεις βίντεο του YouTube και επίσημων, δομημένων κειμένων από δημοσίως διαθέσιμα έγγραφα PDF, κυρίως βιβλία και ακαδημαϊκό υλικό. Για το instruction tuning, μεταφράσαμε υπάρχοντα ξενόγλωσσα σύνολα υψηλής ποιότητας με μία προσαρμοσμένη διαδικασία μετάφρασης, διασφαλίζοντας πολιτισμική συνάφεια και συζητήσεις με διατήρηση του πλαισίου (context) στα Ελληνικά. Κατά τη δημιουργία των δεδομένων, εφαρμόσαμε μια σειρά βημάτων επεξεργασίας, όπως αφαίρεση θορύβου, κανονικοποίηση, φιλτράρισμα με βάση τη γλώσσα και αφαίρεση διπλότυπων, οδηγώντας στην ανάπτυξη ενός αξιόπιστου αγωγού επεξεργασίας.

Τα τελικά σύνολα δεδομένων περιλαμβάνουν πάνω από 2.3 δισεκατομμύρια λέξεις και 6 δισεκατομμύρια tokens, αποτελώντας ένα σημαντικό βήμα προς την εκπαίδευση Ελληνικών LLMs υψηλής ποιότητας. Η εργασία αυτή προσφέρει τόσο επαναχρησιμοποιήσιμη τεχνική υποδομή όσο και επιμελημένα δεδομένα που υποστηρίζουν τη μελλοντική έρευνα και ανάπτυξη στον τομέα του Ελληνικού NLP.

Λέξεις Κλειδιά —Μεγάλα Γλωσσικά Μοντέλα, Γλώσσες Χαμηλών Πόρων, Ελληνικό Σύνολο Δεδομένων, Προεκπαίδευση, Εκπαίδευση Βάσει Οδηγιών, Επιβλεπόμενη Προσαρμογή, Δημιουργία Συνόλων Δεδομένων, Επιμέλεια Δεδομένων, Αγωγός Επεξεργασίας Δεδομένων, Προεπεξεργασία, Μετεπεξεργασία

Abstract

Large Language Models (LLMs) have emerged as powerful tools in Natural Language Processing, propelled by the ever-expanding scale of model sizes and training datasets. While such resources exist for high-resource languages, low-resource languages such as Greek remain significantly underrepresented in modern LLM research and development.

In this thesis, we address this gap by constructing two foundational datasets for Greek LLM development: a pretraining dataset and an instruction tuning dataset. For pretraining, we collected and processed large volumes of conversational data from YouTube transcripts and formal, structured texts from publicly available PDF documents, mostly books and academic material. For instruction tuning, we translated existing high-quality instruction corpora using a custom translation pipeline, ensuring cultural relevance and context-aware conversation in Greek. Throughout the data creation process, we implemented a series of processing steps, including noise removal, formatting normalization, language filtering, and deduplication, leading to the development of a robust processing pipeline.

The final datasets, comprising over 2.3 billion words and 6 billion tokens, mark a significant advancement toward training high-quality Greek LLMs. Our work contributes both reusable infrastructure and curated data to support future research and development in Greek NLP.

Keywords —Large Language Models, Low-Resource Languages, Greek Dataset, Pretraining, Instruction Tuning, Supervised Fine-Tuning, Dataset Creation, Data Curation, Processing Pipeline, Preprocessing, Postprocessing

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους καθηγητές κ. Γεώργιο Στάμου και κ. Μιχάλη Βαζιργιάννη για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσαν να την εκπονήσω στο εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης. Επίσης, ευχαριστώ ιδιαίτερα τον κ. Μιχάλη Βαζιργιάννη για τις ιδέες, την καθοδήγηση και την εμπιστοσύνη που μου έδειξε κατά την διάρκεια της εκπόνησης αυτής της εργασίας.

Ευχαριστώ από καρδιάς τους γονείς μου, Σπύρο και Αλεξία, καθώς και τα αδέρφια μου, Αναστάση, Βασιλική και Μαρία-Χριστίνα, για τη διαρκή στήριξη και την αγάπη τους όλα αυτά τα χρόνια.

Τέλος, ένα ξεχωριστό ευχαριστώ στην κοπέλα μου Ροζαλένα για την κατανόηση και τη στήριξή της σε κάθε μου βήμα.

Contents

List of Figures

List of Tables

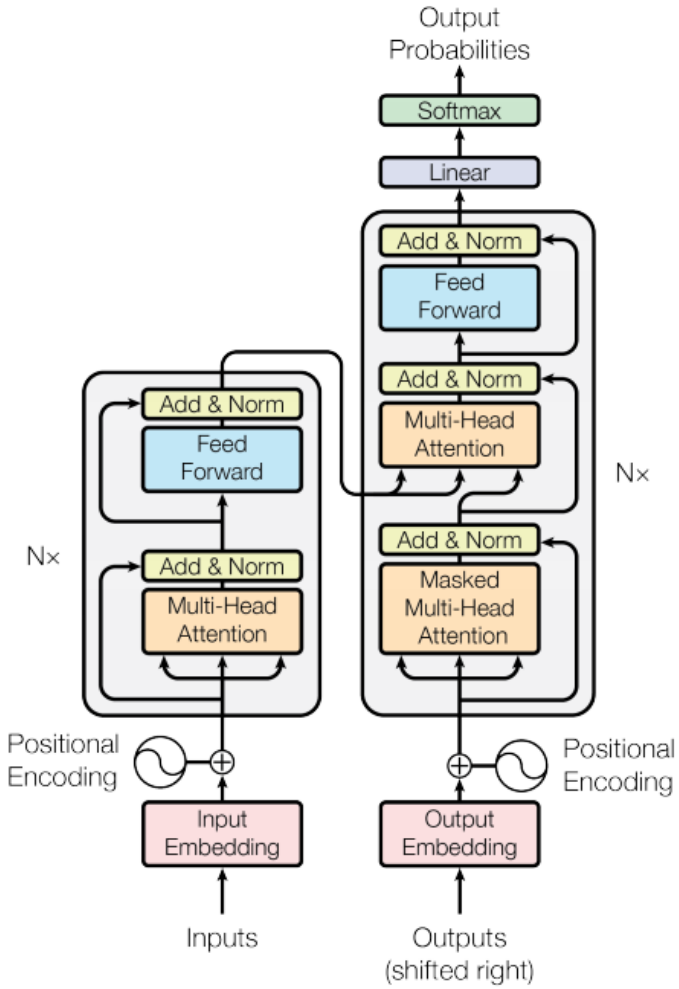
Chapter 0
Εκτεταμένη Περίληψη στα Ελληνικά

Contents

0.1 Θεωρητικό Υπόβαθρο

0.1.1 Μεγάλα Γλωσσικά Μοντέλα (LLMs)

Τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) είναι μοντέλα βαθιάς μάθησης που εκπαιδεύονται σε τεράστιες ποσότητες κειμένου με σκοπό την κατανόηση και παραγωγή φυσικής γλώσσας. Η πλειοψηφία των σύγχρονων LLMs βασίζεται στην **αρχιτεκτονική των Μετασχηματιστών** (Transformers) [1], η οποία εισήγαγε τους **μηχανισμούς αυτοπροσοχής** (self-attention), επιτρέποντας τη μοντελοποίηση μακρινών εξαρτήσεων σε ακολουθίες κειμένου.



Σχήμα 0.1.1: Η αρχιτεκτονική των Μετασχηματιστών, όπως προτάθηκε στο *"Attention Is All You Need"* [1]

Η βασική αρχιτεκτονική των Μετασχηματιστών, όπως φαίνεται στο Σχήμα 0.1.1, περιλαμβάνει δύο συνιστώσες: τον **κωδικοποιητή** (encoder) και τον **αποκωδικοποιητή** (decoder). Στα σύγχρονα μοντέλα, συναντάμε τις εξής διαφορετικές αρχιτεκτονικές παραλλαγές:

- **Μοντέλα μόνο με αποκωδικοποιητή** όπως τα μοντέλα της οικογένειας GPT (Generative Pretrained Transformers)

[2, 3], είναι κατάλληλα για δημιουργικές εργασίες, όπως η δημιουργία κειμένου, η συνομιλία και η παραγωγή κώδικα.

- **Μοντέλα μόνο με κωδικοποιητή** όπως το BERT [4], είναι βελτιστοποιημένα για εργασίες κατανόησης γλώσσας, όπως η ταξινόμηση, η αναγνώριση ονομασμένων οντοτήτων (NER) και η απάντηση σε ερωτήσεις.
- **Μοντέλα με κωδικοποιητή-αποκωδικοποιητή** (seq2seq) όπως τα T5 [5] και BART [6], είναι αποδοτικά σε εργασίες μετάφρασης, σύνοψης και επανεγγραφής κειμένου.

0.1.2 Σύνολα Δεδομένων Προεκπαίδευσης

Η προεκπαίδευση αποτελεί το πρώτο και θεμελιώδες στάδιο στην εκπαίδευση Μεγάλων Γλωσσικών Μοντέλων. Κατά τη διάρκεια της, τα μοντέλα μαθαίνουν γλωσσικά μοτίβα μέσα από τα δεδομένα προεκπαίδευσης, χωρίς την ανάγκη εποπτείας ή χειροκίνητων επισημειώσεων.

Τα σύνολα δεδομένων προεκπαίδευσης αποτελούν, συνεπώς, τα θεμέλια των LLMs, με το μέγεθος, την ποικιλία και την ποιότητά τους να επηρεάζουν την ικανότητα του μοντέλου να γενικεύει και να κατανοεί τη γλώσσα. Επίσης, συνδέονται με την εμφάνιση *αναδυόμενων ικανοτήτων* (emergent abilities), δηλαδή δυνατοτήτων όπως η περίληψη, μετάφραση, συλλογισμός και άλλων, για τις οποίες τα μοντέλα δεν έχουν εκπαιδευτεί [7, 8].

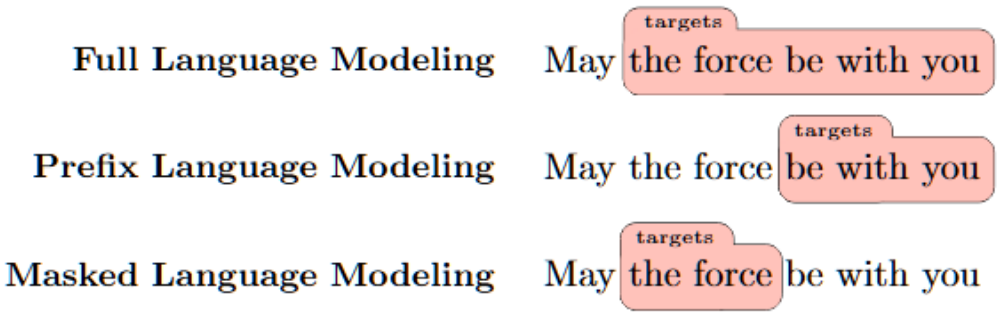
Αποτελούνται από δισεκατομμύρια ή και τρισεκατομμύρια tokens, τα οποία συλλέγονται από ποικιλία πηγών, όπως ιστοσελίδες, βιβλία, εγκυκλοπαίδειες, ακαδημαϊκά έγγραφα και άλλα [9].

Η ποιότητα των δεδομένων είναι εξίσου σημαντική με την ποσότητά τους [10]. Η εμφάνιση θορύβου, επαναλαμβανόμενων δεδομένων ή δεδομένων σε γλώσσα διαφορετική από αυτήν που προορίζεται για την εκπαίδευση, μπορεί να επηρεάσει αρνητικά την εκπαίδευση και κατά συνέπεια την επίδοση του μοντέλου [11, 12]. Για το λόγο αυτό, είναι απαραίτητη η εφαρμογή αυστηρών βημάτων επεξεργασίας όπως:

- **Καθαρισμός και κανονικοποίηση μορφοποίησης**
- **Ανίχνευση γλώσσας και φιλτράρισμα**
- **Αφαίρεση διπλότυπων εγγράφων**

Τέλος, η επιλογή του προεκπαιδευτικού στόχου έχει σημαντικό αντίκτυπο στην χρηστικότητα του μοντέλου, αφού καθορίζει τον τρόπο με τον οποίο μαθαίνει αναπαραστάσεις της γλώσσας [13, 14]. Η απόφαση αυτή εξαρτάται άμεσα και από την επιλεγμένη αρχιτεκτονική του μοντέλου, δηλαδή:

- **Αιτιώδης ή Πλήρης Γλωσσική Μοντελοποίηση (CLM/FLM):** πρόβλεψη του επόμενου token με βάση μόνο τα προηγούμενα - χρησιμοποιείται σε αποκωδικοποιητές.
- **Μοντελοποίηση Γλώσσας με Πρόθεμα (PLM):** Για να μπορούν τα μοντέλα κωδικοποιητή-αποκωδικοποιητή να εκτελούν μοντελοποίηση γλώσσας, πρέπει να οριστεί ένα πρόθεμα όπου το μοντέλο δεν περιορίζεται σε αιτιώδη (causal) προσοχή σε αυτό. Το μοντέλο προβλέπει κάθε token εκτός του προθέματος, δεδομένων όλων των προηγούμενων διακριτικών.
- **Γλωσσική Μοντελοποίηση με Μάσκα (MLM):** πρόβλεψη των κρυμμένων tokens με βάση τα συμφραζόμενα (αμφίδρομα) - χρησιμοποιείται σε κωδικοποιητές.



Σχήμα 0.1.2: Παραδείγματα εισόδου και στόχου πρόβλεψης για προεκπαιδευτικό στόχο FLM, PLM και MLM. [14]

0.1.3 Σύνολα Δεδομένων για Instruction Tuning

Μετά την ολοκλήρωση του σταδίου της προεκπαίδευσης, τα LLMs αποκτούν επαρκή γνώση για την παραγωγή κειμένου. Ωστόσο, για να ανταποκρίνονται με συνέπεια σε ρητές οδηγίες χρηστών, απαιτείται περαιτέρω ευθυγράμμιση (alignment). Η ευθυγράμμιση αυτή επιτυγχάνεται μέσω της **εκπαίδευσης Βάσει Οδηγίων** (Instruction Tuning) ή αλλιώς **Επιβλεπόμενης Προσαρμογής** (Supervised Fine-Tuning - SFT), κατά την οποία τα μοντέλα εκπαιδεύονται σε ζεύγη (οδηγία, απάντηση) ή διαλόγους πολλαπλών γύρων [15].

Σύμφωνα με μελέτες [16], η μέθοδος αυτή έχει αποδειχθεί ότι βοηθάει τα μοντέλα να γενικεύουν καλύτερα, ενώ ταυτόχρονα βελτιώνει την ευχρηστία και την ακρίβειά τους σε άγνωστες εργασίες (zero-shot).

Στην πράξη, τέτοια σύνολα δεδομένων μπορούν να δημιουργηθούν μέσω:

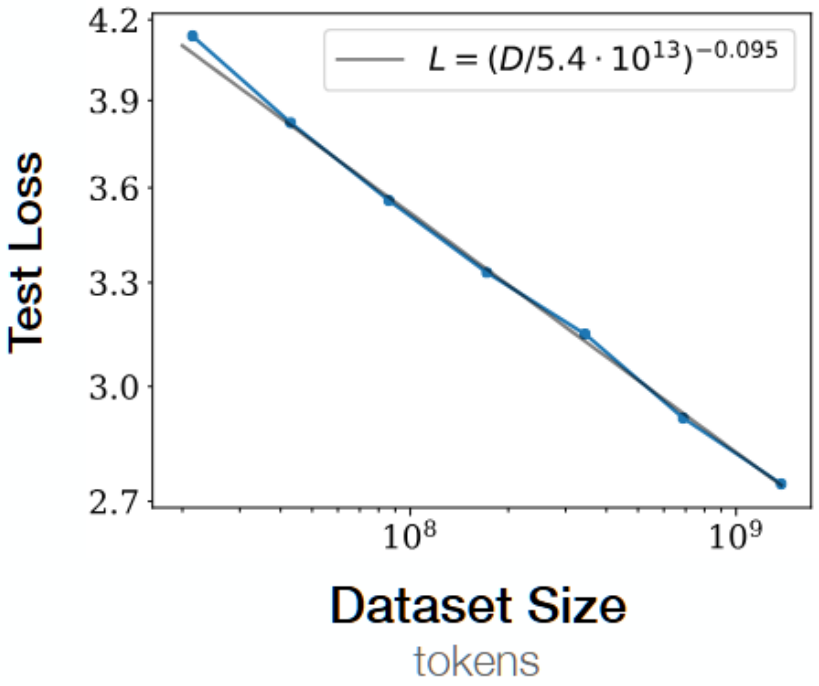
- Ανθρώπινων επισημειώσεων
- Αυτόματης δημιουργίας με χρήση LLMs [15, 17]
- Μετάφρασης υπαρχόντων αγγλόφωνων συνόλων δεδομένων [18, 19]

0.1.4 Προκλήσεις για τις Γλώσσες Χαμηλών Πόρων

Η απόδοση των LLMs εξαρτάται σε μεγάλο βαθμό από το μέγεθος και την ποιότητα των δεδομένων. Αυτό αποτυπώνεται στους νόμους κλιμάκωσης [20], οι οποίοι αναφέρονται σε εμπειρικές σχέσεις που δείχνουν ότι η απόδοση και οι δυνατότητες γενίκευσης των μοντέλων βελτιώνονται με προβλέψιμο τρόπο με την αύξηση του μεγέθους των δεδομένων, των παραμέτρων του μοντέλου και των πόρων εκπαίδευσης.

Πράγματι, η ανάπτυξη Μεγάλων Γλωσσικών Μοντέλων βασίζεται κατά κύριο λόγο στη διαθεσιμότητα τεράστιων ποσοτήτων δεδομένων από διάφορες πηγές, κυρίως στα Αγγλικά. Κατά συνέπεια, οι λεγόμενες *γλώσσες χαμηλών πόρων* (low-resource languages), όπως τα Ελληνικά, παραμένουν σημαντικά υποεκπροσωπούμενες στα περισσότερα σύνολα δεδομένων μεγάλης κλίμακας.

Η ελληνική γλώσσα, αν και πλούσια σε μορφολογία και πολιτισμική σημασία, πάσχει από περιορισμένη παρουσία στο διαδίκτυο και από την απουσία αποδοτικών tokenizer. Οι παράγοντες αυτοί καθιστούν πολύπλοκη την ανάπτυξη Μεγάλων Γλωσσικών Μοντέλων υψηλής ποιότητας στα Ελληνικά.



Σχήμα 0.1.3: Η απόδοση των Γλωσσικών Μοντέλων βελτιώνεται για μεγαλύτερα σύνολα δεδομένων, "Scaling Laws for Neural Language Models" [20]

Ακόμη και σε πρόσφατες διεθνείς προσπάθειες για τη συλλογή και δημιουργία μεγάλων πολυγλωσσικών δεδομένων, όπως τα **CulturaX** [21], **mC4** [22], **TeuKen-7B** [11, 23] και **HPLT** [24], η ελληνική γλώσσα καλύπτεται περιορισμένα, τόσο σε όγκο όσο και σε ποικιλία θεματολογίας.

Για την αντιμετώπιση αυτών των προκλήσεων απαιτούνται στοχευμένες προσπάθειες συλλογής, καθαρισμού και επιμέλειας ελληνικών δεδομένων, κάτι που αποτελεί βασικό στόχο της παρούσας εργασίας.

Αξιοσημείωτες πρόσφατες προσπάθειες για τη δημιουργία ελληνικών LLMs είναι το **Meltemi** [18] και το **LLaMA-Krikri** [25].

0.1.5 Βασικές Έννοιες

Στην παρούσα ενότητα γίνεται συνοπτική αναφορά και επεξήγηση των εργαλείων και τεχνικών που χρησιμοποιήθηκαν κατά τη συλλογή και την επεξεργασία των δεδομένων:

- **Συστήματα Αναγνώρισης Ομιλίας (ASR):** Επιτρέπουν τη μετατροπή ακουστικών δεδομένων σε γραπτό κείμενο. Τα σύγχρονα μοντέλα είναι κατά βάση βαθιά νευρωνικά δίκτυα, και η λειτουργία τους περιλαμβάνει συνήθως τα ακόλουθα στάδια:
    1. **Προεπεξεργασία:** αφαίρεση θορύβου από το ηχητικό σήμα.
    2. **Νευρωνικό Δίκτυο:** (CNN, RNN, Transformers, Conformers) για την αντιστοίχιση φωνητικών μονάδων σε κείμενο [26].

3. **Γλωσσικό Μοντέλο:** για τη διόρθωση και πρόβλεψη των πιο πιθανών λέξεων ή φράσεων.

- **Οπτική Αναγνώριση Χαρακτήρων (OCR):** Η μέθοδος εξαγωγής κειμένου από σαρωμένα έγγραφα ή φωτογραφίες. Συνήθως χρησιμοποιούν συνδυασμό Συνελικτικών (CNN) και Αναδρομικών Νευρωνικών Δικτύων (RNN), ή αρχιτεκτονικές βασισμένες στους Μετασχηματιστές.

- **Αναγνώριση Γλώσσας (LID):** Η αυτόματη αναγνώριση της γλώσσας ενός κειμένου είναι απαραίτητη για την αφαίρεση κειμένων που δεν περιλαμβάνονται στο λεξιλόγιο του μοντέλου προς εκπαίδευση [11]. Συνήθως χρησιμοποιούν στατιστικές μεθόδους, όπως το Bag-of-n-grams [27], ή αρχιτεκτονικές βασισμένες στους Μετασχηματιστές.

- **Αφαίρεση Διπλότυπων:** Η διαδικασία αφαίρεσης διπλότυπων (deduplication) είναι κρίσιμη για τη διασφάλιση της ποιότητας των δεδομένων. Πολλές μελέτες έχουν αναλύσει την ανάγκη για αφαίρεση διπλότυπων κειμένων, εστιάζοντας αρχικά στην αποδοτικότητα των υπολογιστικών πόρων, αναφέροντας ότι τέτοιες τεχνικές επιτρέπουν τη γρηγορότερη σύγκλιση των μοντέλων πετυχαίνοντας ίδια ή και καλύτερα αποτελέσματα [28].

  Το σημαντικότερο πρόβλημα στην ύπαρξη διπλότυπων κατά την εκπαίδευση ενός LLM είναι ότι προκαλούν υπερπροσαρμογή (overfitting) σε επαναλαμβανόμενα μοτίβα, οδηγούν τα μοντέλα σε εμφάνιση προκαταλήψεων, μειώνουν την ποικιλομορφία των γλωσσικών δομών που μαθαίνει το μοντέλο και γενικώς βλάπτουν τις εσωτερικές δομές προσοχής (attention) των μοντέλων, οδηγώντας τα σε απομνημόνευση (memorization) αντί της γενίκευσης [20, 29].

  Χρησιμοποιήθηκε ο αλγόριθμος MinHash [30, 31] σε συνδυασμό με Locality Sensitive Hashing (LSH) [32], ώστε να ανιχνεύονται και να αφαιρούνται έγγραφα με υψηλή ομοιότητα, χωρίς ανάγκη για πλήρη σύγκριση. Η μέθοδος αυτή είναι ευρέως χρησιμοποιούμενη για τον εντοπισμό όμοιων εγγράφων [33], ενώ έχει αποδειχθεί πειραματικά ότι προσφέρει την καλύτερη ισορροπία μεταξύ ακρίβειας, ανάκλησης και υπολογιστικής αποδοτικότητας [34].

0.2 Δημιουργία Συνόλων Δεδομένων Προεκπαίδευσης

Ένας από τους βασικούς στόχους της παρούσας εργασίας είναι η δημιουργία ενός συνόλου δεδομένων προεκπαίδευσης μεγάλης κλίμακας για την ελληνική γλώσσα.

Για τη δημιουργία ενός τέτοιου συνόλου, δόθηκε έμφαση τόσο στην ποικιλομορφία και τον όγκο των δεδομένων, όσο και στην επεξεργασία και το φιλτράρισμά τους, ώστε να διασφαλιστεί η ποιότητα και η γλωσσική καταλληλότητα των δεδομένων. Οι δύο κύριες πηγές που χρησιμοποιήθηκαν είναι:

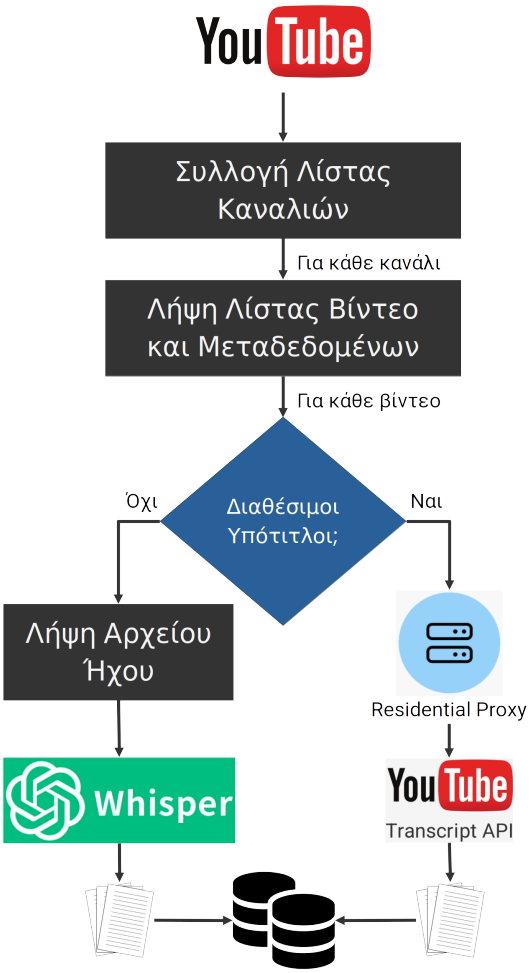
- **Απομαγνητοφωνήσεις από βίντεο YouTube**
- **Έγγραφα PDF από δημόσια αποθετήρια**

0.2.1 Συλλογή και Επεξεργασία Δεδομένων από το YouTube

Το **YouTube** αποτελεί ένα μεγάλο αποθετήριο περιεχομένου, το οποίο περιλαμβάνει μεγάλους όγκους συνομιλιακών δεδομένων σε καθημερινή, ανεπίσημη γλώσσα. Στην περίπτωση της ελληνικής γλώσσας, αποτελεί μία από τις λίγες

μεγάλης κλίμακας, δημόσια διαθέσιμες πηγές που παρέχουν πρόσβαση σε σύγχρονο, πραγματικό διάλογο σε ένα ευρύ φάσμα θεματολογίας.

Για τη συλλογή των απομαγνητοφωνήσεων βίντεο, σχεδιάστηκε και υλοποιήθηκε ένας αγωγός πολλών σταδίων, όπως φαίνεται στο Σχήμα 0.2.1.



Σχήμα 0.2.1: Αγωγός συλλογής απομαγνητοφωνήσεων από το YouTube

Το πρώτο βήμα αφορά την **επιλογή καναλιών**. Αυτή πραγματοποιήθηκε σε μεγάλο βαθμό αυτόματα, με τεχνικές απόξεσης ιστού (web-scraping) για την εύρεση των δημοφιλέστερων ελληνικών καναλιών ανά κατηγορία και θεματολογία περιεχομένου. Η λίστα αυτή στη συνέχεια ελέγχθηκε χειροκίνητα για την αφαίρεση ξενόγλωσσων και κατηργημένων καναλιών, και εμπλουτίστηκε περαιτέρω με επιλεγμένα κανάλια και λίστες αναπαραγωγής, δίνοντας έμφαση σε βίντεο και podcasts υψηλής ποιότητας περιεχομένου, με σωστή άρθρωση και συνέπεια στην ελληνική γλώσσα. Η αρχική συλλογή περιλαμβάνει **238 κανάλια και 65 λίστες αναπαραγωγής**, με ένα σύνολο από **778,998 βίντεο**.

Στη συνέχεια, για τη **συλλογή των απομαγνητοφωνήσεων (transcripts)** των παραπάνω βίντεο, χρησιμοποιήθηκε η βιβλιοθήκη **YouTube Transcript API** [35]. Από τη διαδικασία αυτή, εξήχθησαν απομαγνητοφωνήσεις για το 21% των συλλεχθέντων βίντεο της αρχικής συλλογής, καθώς τα περισσότερα βίντεο δεν διέθεταν υπότιτλους. Για

την περαιτέρω ενίσχυση του παραπάνω συνόλου δεδομένων, χρησιμοποιήθηκε επιπλέον το **Σύστημα Αναγνώρισης Ομιλίας (ASR) Whisper** της OpenAI [36], λόγω της υψηλής απόδοσής του στα Ελληνικά.

Για να εξισορροπηθούν οι επιδόσεις και η αποδοτικότητα, χρησιμοποιήθηκε μια κβαντισμένη έκδοση του **Whisper-Large-v2**, με τη βοήθεια της βιβλιοθήκης **faster-whisper** [37]. Αυτή η λύση παρείχε την καλύτερη ισορροπία μεταξύ ακρίβειας και υπολογιστικού κόστους. Λόγω των απαιτήσεων σε υπολογιστικούς πόρους, οι απομαγνητοφωνήσεις περιορίστηκαν σε ένα μικρό ποσοστό επιλεγμένων βίντεο υψηλής ποιότητας από τις ακόλουθες κατηγορίες: Λίστες αναπαραγωγής, Χειροκίνητα επιλεγμένα, Ειδήσεις και Ιστορικά, συλλέγοντας επιπλέον **983 απομαγνητοφωνήσεις βίντεο**.

Οι απομαγνητοφωνήσεις που εξήχθησαν από την παραπάνω διαδικασία ήταν σε μεγάλο βαθμό θορυβώδεις, παρουσιάζοντας συνήθη σφάλματα που προκύπτουν από συστήματα ASR αλλά και από τη φύση του περιεχομένου, όπως:

- Μερική ή ολική απώλεια σημείων στίξης
- Έλλειψη κεφαλαίων
- Ορθογραφικά λάθη
- Λέξεις δισταγμού (π.χ. εεε, χμμ)
- Εισαγωγές καναλιών και διαφημίσεις

Η παρουσία τέτοιων σφαλμάτων μειώνει σημαντικά την ποιότητα του συνόλου και καθιστά τις απομαγνητοφωνήσεις ακατάλληλες για χρήση σε προεκπαίδευση Μεγάλων Γλωσσικών Μοντέλων, λόγω της επιρροής της ποιότητας των δεδομένων στην απόδοση των μοντέλων [5, 38]. Για την εξασφάλιση της ποιότητας, εστιάσαμε σε τρεις ξεχωριστές εργασίες για τον **καθαρισμό των εξαγόμενων κειμένων**: **αφαίρεση θορύβου**, **κανονικοποίηση κειμένου** και **επαναφορά σημείων στίξης και κεφαλαίων**.

Αρχικά, εξετάστηκε και αξιολογήθηκε η χρήση εργαλείων ανοιχτού κώδικα (open-source) για Επεξεργασία Φυσικής Γλώσσας (NLP) στα ελληνικά, και συγκεκριμένα τα **GR-NLP-TOOLKIT** [39], **Greek BART** (seq2seq) [40] και **Greek BERT** [41].

Τα μοντέλα αυτά παρέχουν βασικές λειτουργικότητες, αλλά δεν είναι σε θέση να υποστηρίξουν την πολυσταδιακή επεξεργασία που απαιτείται. Συνεπώς, εξετάσαμε τη **χρήση Μεγάλων Γλωσσικών Μοντέλων για τη διόρθωση** των σφαλμάτων, ορμώμενοι από τις εκτενείς μελέτες των τελευταίων ετών που επιβεβαιώνουν την αποδοτικότητα των LLMs σε πολλές εργασίες NLP [42--44].

Έτσι, με την κατάλληλη καθοδήγηση (prompting), χρησιμοποιήθηκε το μοντέλο **gpt-4o-mini** της OpenAI για τη διόρθωση των προαναφερθέντων σφαλμάτων.

Τέλος, για τη διασφάλιση της **γλωσσικής καταλληλότητας**, υλοποιήθηκε μία διαδικασία δύο σταδίων για την ανίχνευση και απομάκρυνση απομαγνητοφωνήσεων που δεν αναγνωρίζονταν με υψηλή βεβαιότητα ως ελληνικές, χρησιμοποιώντας τα μοντέλα **fastText** [27, 45] (1ο στάδιο) και **GlotLID** [46] (2ο στάδιο).

0.2.2 Εξαγωγή και Επεξεργασία Εγγράφων PDF

Σε αντίθεση με τον προφορικό λόγο, τα γραπτά έγγραφα προσφέρουν πρόσβαση σε επίσημο, δομημένο λόγο, ο οποίος είναι απαραίτητος για την εκπαίδευση γλωσσικών μοντέλων γενικής χρήσης. Για να διασφαλιστεί ένα αντιπροσωπευ-

τικό δείγμα γραπτής ελληνικής γλώσσας σε ποικιλία τομέων, συλλέχθηκε υλικό από τις ακόλουθες πηγές:

1. Εθνικό Αρχείο Διδακτορικών Διατριβών: Μία πηγή από διδακτορικές διατριβές οι οποίες εκπονήθηκαν στα ελληνικά πανεπιστήμια, καθώς και αυτές που εκπονήθηκαν από Έλληνες σε πανεπιστήμια του εξωτερικού και αναγνωρίστηκαν από τον αρμόδιο εθνικό φορέα, τον ΔΟΑΤΑΠ.
2. ebooks.edu.gr: Ο επίσημος ψηφιακός χώρος του Υπουργείου Παιδείας, Θρησκευμάτων & Αθλητισμού (ΥΠΑΙΘΑ) για τη διάθεση των ψηφιακών μορφών των σχολικών βιβλίων. Περιέχει όλα τα σχολικά βιβλία για το Δημοτικό, το Γυμνάσιο, το Γενικό και το Επαγγελματικό Λύκειο (ΕΠΑ.Λ) σε διάφορες ψηφιακές μορφές.
3. OpenBook, free-ebooks, eBooks4Greeks: Ψηφιακά αποθετήρια με χιλιάδες ελληνικά ψηφιακά βιβλία που είτε είναι δημόσιας χρήσης είτε διανέμονται ελεύθερα και νόμιμα στο διαδίκτυο από τους δημιουργούς ή τους εκδότες τους.

Η **εξαγωγή των κειμένων** από τα συλλεγμένα αρχεία πραγματοποιήθηκε κατά κύριο λόγο με τη χρήση της βιβλιοθήκης **PyMuPDF** [47], με τη χρήση της οποίας εξήχθησαν τα καθαρά περιεχόμενα από την πλειοψηφία των παραπάνω εγγράφων PDF, στα οποία το κείμενο είναι ενσωματωμένο στο αρχείο.

Στα υπόλοιπα αρχεία, για τα οποία η απευθείας εξαγωγή κειμένου δεν ήταν εφικτή, το κείμενο δεν ήταν ενσωματωμένο, με αποτέλεσμα οι επιμέρους σελίδες να αντιμετωπίζονται ως εικόνες, πιθανώς προερχόμενες από σαρωμένα έγγραφα. Για το μικρό αυτό υποσύνολο εφαρμόστηκαν τεχνικές **Οπτικής Αναγνώρισης Χαρακτήρων (OCR)**, ώστε να μετατραπεί το περιεχόμενο των αρχείων σε επεξεργάσιμη μορφή κειμένου. Συνολικά, το τελικό υποσύνολο των εγγράφων που απαιτούσαν OCR ήταν μικρότερο του 3% του συνόλου.

Για την εξασφάλιση της ποιότητας των εξαχθέντων κειμένων, εφαρμόστηκαν τα παρακάτω βήματα **επεξεργασίας και καθαρισμού**:

1. **Κανονικοποίηση Διάταξης και Μορφοποίησης:** Αφορά την αφαίρεση θορύβου που οφείλεται στη μορφή των βιβλίων, και περιλαμβάνει την εμφάνιση επαναλαμβανόμενων κεφαλίδων και υποσέλιδων, αριθμό σελίδας, κ.ά.
2. **Εντοπισμός Γλώσσας και Φιλτράρισμα:** Για τη διασφάλιση της γλωσσικής συνέπειας σε όλο το σύνολο δεδομένων, χρησιμοποιώντας τη διαδικασία δύο σταδίων με τα μοντέλα **fastText** και **GlotLID**, όπως περιγράφηκε στην προηγούμενη ενότητα.
3. **Αφαίρεση Διπλότυπων:** Χρησιμοποιώντας τον αλγόριθμο MinHash LSH μέσω της υλοποίησης της βιβλιοθήκης **text-dedup** [48], ώστε να εντοπίσουμε και να αφαιρέσουμε έγγραφα υψηλής ομοιότητας ή ταυτόσημα, τα οποία εμφανίζονταν σε περισσότερες από μία πηγές ή περιείχαν επαναλήψεις περιεχομένου.

0.3 Δημιουργία Συνόλου Instruction Tuning

Η εκπαίδευση Μεγάλων Γλωσσικών Μοντέλων βάσει οδηγιών (instruction tuning) παίζει καθοριστικό ρόλο στην ευθυγράμμισή τους με τις προθέσεις των χρηστών, επιτρέποντάς τους να μεταβούν από την γενική πρόβλεψη επόμενης λέξης στην τήρηση και ακολούθηση ανθρώπινων οδηγιών. Με τον τρόπο αυτό, βελτιώνεται η απόδοση των

μοντέλων σε ένα ευρύ φάσμα εργασιών και ενισχύεται η ικανότητά τους να αλληλεπιδρούν με ασφάλεια με τους χρήστες [49]

Δεδομένου ότι δεν υπάρχουν διαθέσιμα δεδομένα μεγάλου όγκου αυτού του τύπου για την ελληνική γλώσσα, επιλέχθηκε η στρατηγική της **μετάφρασης υπαρχόντων συνόλων δεδομένων**. Η προσέγγιση αυτή έχει χρησιμοποιηθεί εκτενώς σε συναφείς προσπάθειες εκπαίδευσης LLMs σε γλώσσες χαμηλών πόρων [18, 19, 25]. Ως βάση, χρησιμοποιήθηκαν τα ακόλουθα δύο σύνολα υψηλής ποιότητας:

- **WildChat:** Αποτελείται από 1 εκατομμύριο διαλόγους πολλαπλών γύρων μεταξύ χρηστών και chat-bots. Περιέχει ένα ευρύ φάσμα αλληλεπιδράσεων συμπεριλαμβανομένων διφορούμενων αιτημάτων χρηστών, εναλλαγή κώδικα και θεμάτων, πολιτικές συζητήσεις κ.λπ. [50].
- **UltraChat:** Συνθετικό σύνολο δεδομένων με διαλόγους πολλαπλών γύρων, δημιουργημένων από την προσομοίωση συνομολίας μεταξύ δύο πρακτόρων ChatGPT, όπου ο ένας παίζει το ρόλο του χρήστη για τη δημιουργία ερωτήσεων, ενώ ο άλλος παράγει τις αντίστοιχες απαντήσεις. Η κύρια αρχή του συνόλου είναι να αυξήσει την ποικιλομορφία των δεδομένων, σε αντίθεση με τα περισσότερα σύνολα δεδομένων που εστιάζουν σε συγκεκριμένες εργασίες όπως απαντήσεις σε ερωτήσεις, επανεγγραφή και σύνοψη. [51]

Αρχικά εξετάστηκαν δύο από τα δημοφιλέστερα μοντέλα μηχανικής μετάφρασης: το **Helsinki-NLP Opus-MT** [52] και το **NLLB-200** (No Language Left Behind) της Meta [52]. Παρότι αποδίδουν ικανοποιητικά σε απλές μεταφράσεις επιπέδου πρότασης, περιορίζονται σημαντικά σε συνομιλιακό περιεχόμενο μεγάλης έκτασης, κυρίως λόγω των μικρών παραθύρων συμφραζομένων (context windows) με 512 και 1024 tokens αντίστοιχα. Η ανάγκη τεμαχισμού οδηγεί σε απώλεια συνοχής και λάθη στην εναλλαγή ρόλων, καθιστώντας τα ακατάλληλα για υψηλής ποιότητας μεταφράσεις διαλόγου.

Αντιθέτως, τα σύγχρονα Μεγάλα Γλωσσικά Μοντέλα υποστηρίζουν ευρύ παράθυρο συμφραζομένων και η ικανότητά τους έχει μελετηθεί ιδιαιτέρως λόγω της σημαντικότητας της μηχανικής μετάφρασης, ως μία απ'τις θεμελιώδεις εργασίες της Επεξεργασίας Φυσικής Γλώσσας (NLP) [53]. Με βάση τη σύγχρονη έρευνα, τα LLMs είναι εξαιρετικά ικανά σε μεταφράσεις μεγάλων εκτάσεων κειμένου [54], ξεπερνώντας την ποιότητα δημοφιλών υπηρεσιών μετάφρασης [55].

Για την εργασία της μετάφρασης χρησιμοποιήθηκε το μοντέλο **Claude Sonnet** της Anthropic, μέσω της πλατφόρμας **Amazon Bedrock**. Το μοντέλο επιλέχθηκε λόγω της ισχυρής απόδοσής του σε πολυγλωσσικές εργασίες και της υψηλής ακρίβειας στις μεταφράσεις [56].

Πριν τη μετάφραση, εφαρμόστηκε **προεπεξεργασία** για τον καθαρισμό των αρχικών συνόλων. Συγκεκριμένα, αφαιρέθηκαν διαλόγοι με:

- Κενά ή μη φυσικά μηνύματα
- Διπλότυπα
- Περιεχόμενο σε μη υποστηριζόμενες γλώσσες

Έπειτα, τροποποιήθηκαν οι διάλογοι ώστε να ακολουθηθεί κοινή μορφοποίηση στα σύνολα δεδομένων, μετατρέποντάς τους στη μορφή **Chat-ML** η οποία χρησιμοποιείται ευρέως και δομεί τα δεδομένα ως μία λίστα από μηνύματα που χαρακτηρίζονται από το ρόλο του "συντάκτη" του μηνύματος, ως εξής:

```
[
    {"role": "user", "content": "Prompt 1"},
    {"role": "assistant", "content": "Answer 1."},
    {"role": "user", "content": "Prompt 2"},
    {"role": "assistant", "content": "Answer 2."},
    ...
]
```

Για την καθοδήγηση της μετάφρασης, σχεδιάστηκαν εκτενή prompts με σαφείς οδηγίες για:

- Διατήρηση του αρχικού νοήματος
- Παράλειψη μη μεταφράσιμων στοιχείων (π.χ. snippets κώδικα)
- Αποφυγή ενεργής απάντησης από το μοντέλο
- **Πολιτισμική προσαρμογή**: μετάφραση φράσεων, ιδιωμάτων ή παραδειγμάτων στο ελληνικό πολιτισμικό πλαίσιο

Ο στόχος της εκτενούς περιγραφής ήταν να αποσαφηνιστεί ο ρόλος του μοντέλου και η αναμενόμενη μορφή του παραγόμενου αποτελέσματος, αποσκοπώντας παράλληλα στην αποτροπή πιθανών σφαλμάτων που παρατηρήθηκαν κατά τη φάση των αρχικών δοκιμών, όπως η προσπάθεια μετάφρασης αποσπασμάτων κώδικα ή άλλων μη μεταφράσιμων στοιχείων, η ακούσια απάντηση σε ερωτήσεις του κειμένου, και παραβιάσεις της ζητούμενης συμπεριφοράς.

Μετά την παραγωγή των μεταφρασμένων συνόλων, ακολούθησε **μετεπεξεργασία**, κατά την οποία:

- Εντοπίστηκαν και αφαιρέθηκαν διαλόγοι με μη φυσική χρήση της ελληνικής
- Απορρίφθηκαν περιπτώσεις με ασυνέπεια ρόλων ή αδύναμες απαντήσεις

0.4 Στατιστικά και Ανάλυση Συνόλων Δεδομένων

Η παρούσα ενότητα συνοψίζει τα βασικά χαρακτηριστικά των τελικών συνόλων δεδομένων, τόσο για την προεκπαίδευση όσο και για την εκπαίδευση βάσει οδηγιών. Παρουσιάζονται στατιστικά μεγέθη όπως αριθμός εγγράφων, λέξεων και tokens, καθώς και η επίδραση των βημάτων καθαρισμού στην τελική ποσότητα και ποιότητα των δεδομένων.

0.4.1 Σύνολα Δεδομένων Προεκπαίδευσης

Τα δεδομένα προεκπαίδευσης αποτελούνται από δύο κύριες πηγές: **απομαγνητοφωνήσεις YouTube** και **έγγραφα PDF**.

Η διαδικασία συλλογής δεδομένων από το **YouTube** περιλάμβανε αρχικά **778,998 βίντεο** από **238 κανάλια** και **65 λίστες αναπαραγωγής**, κατηγοριοποιημένα σε **18 θεματικές ενότητες** με βάση το περιεχόμενό τους. Η διαδικασία

λήψης υποτίτλων σε συνδυασμό με τις αυτόματες απομαγνητοφωνήσεις μέσω συστημάτων ASR είχε ως αποτέλεσμα την εξαγωγή **165,758 αρχείων απομαγνητοφωνήσεων**, δηλαδή περίπου 21% της αρχικής συλλογής.

Η εφαρμογή των βημάτων προεπεξεργασίας για κανονικοποίηση, επαναφορά σημείων στίξης, διορθώσεις και αφαίρεση θορυβώδους ή μη ελληνικού περιεχομένου, είχε ως αποτέλεσμα τη δημιουργία ενός υψηλής ποιότητας συνόλου καθημερινού, συνομιλιακού λόγου. Τα τελικά στατιστικά παρουσιάζονται στον Πίνακα 0.4.1:

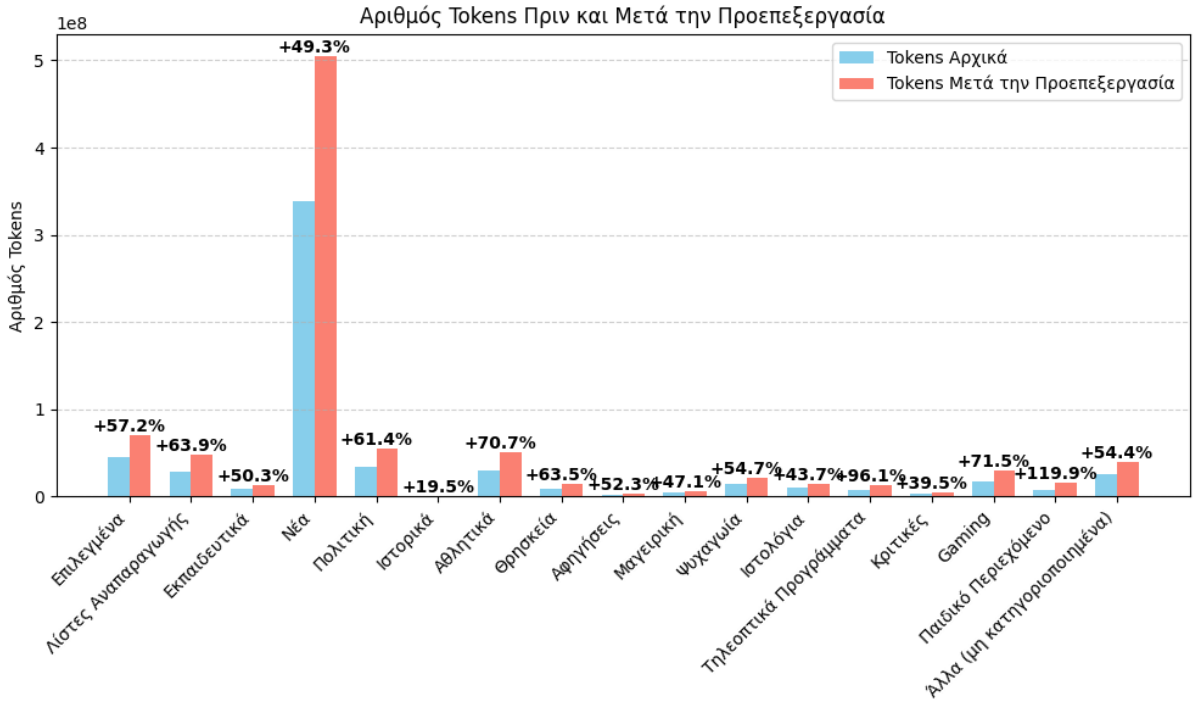
Πίνακας 0.4.1: YouTube: Στατιστικά Προεπεξεργασμένων Απομαγνητοφωνήσεων

Κατηγορία	# Απομαγνητοφωνήσεων	Πλήθος Λέξεων	Πλήθος Tokens
Επιλεγμένα	8,479	32,406,415	70,773,827
Λίστες Αναπαραγωγής	4,646	21,650,973	47,479,107
Εκπαιδευτικά	1,838	6,145,491	13,313,326
Νέα	108,701	226,853,323	505,022,794
Πολιτική	2,472	25,313,555	55,532,496
Ιστορικά	76	67,788	165,836
Αθλητικά	5,763	22,590,427	51,020,406
Θρησκεία	1,307	6,460,612	14,273,951
Αφηγήσεις	623	1,572,275	3,342,010
Μαγειρική	1,707	2,955,733	6,736,421
Ψυχαγωγία	4,567	9,943,517	22,087,924
Ιστολόγια	3,207	6,389,353	13,876,388
Τηλεοπτικά Προγράμματα	1,873	6,161,805	13,543,198
Κριτικές	834	2,167,610	4,418,575
Gaming	3,174	13,211,476	29,214,264
Παιδικό Περιεχόμενο	550	6,293,984	16,116,640
Μουσική	0	0	0
Άλλα (μη κατηγοριοποιημένα)	5,536	17,980,533	39,555,767
Σύνολο	**155,353**	**408,164,870**	**906,472,930**

Αξίζει να σημειωθεί ότι, παρά τη συνολική μείωση του πλήθους των βίντεο, ο αριθμός των tokens στο τελικό υποσύνολο είναι αυξημένος κατά 54% (Σχήμα 0.4.1), γεγονός που οφείλεται κυρίως στην ανασύνθεση προτάσεων, την αποκατάσταση στίξης και τη διόρθωση γλωσσικών λαθών.

Όσον αφορά τα **έγγραφα PDF**, η αρχική συλλογή περιλάμβανε **25,214 έγγραφα** από **5 διαφορετικές πηγές**. Η εξαγωγή κειμένου ήταν εφικτή για το μεγαλύτερο ποσοστό των περιπτώσεων, διαμορφώνοντας ένα αρχικό σύνολο **24,469 αρχείων κειμένου**.

Η εφαρμογή των βημάτων προεπεξεργασίας για κανονικοποίηση, αφαίρεση μη ελληνικού περιεχομένου και διπλότυπων εγγράφων, οδήγησε στον καθαρισμό του συνόλου με το τελικό σύνολο να περιλαμβάνει **20,965 έγγραφα**. Τα συγκεντρωτικά στατιστικά παρουσιάζονται στον Πίνακα 0.4.2:



Σχήμα 0.4.1: Απομαγνητοφωνήσεις YouTube: Αριθμός Tokens πριν και μετά την Προεπεξεργασία

Πίνακας 0.4.2: Έγγραφα PDF: Στατιστικά Προεπεξεργασμένων Κειμένων

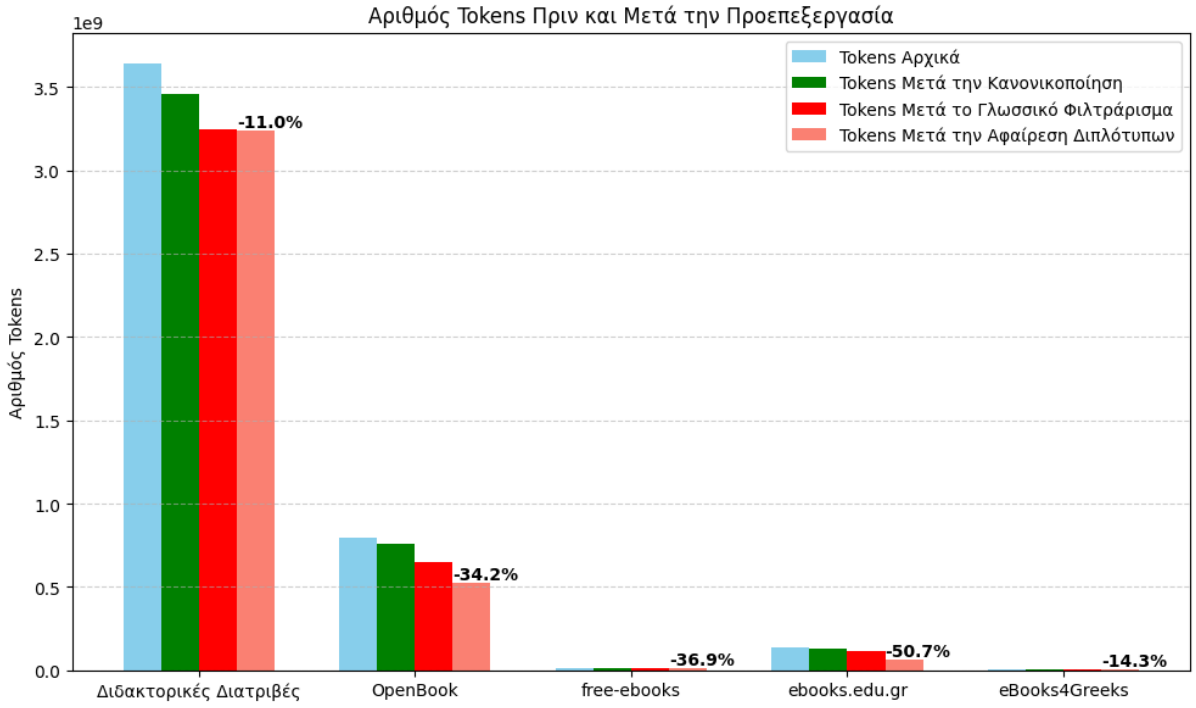
Πηγή	# Εγγράφων	Πλήθος Λέξεων	Πλήθος Tokens
Διδακτορικές Διατριβές	14,889	1,175,320,836	3,240,554,150
OpenBook	5,245	194,908,100	522,816,594
free-ebooks	207	3,725,473	9,689,594
ebooks.edu.gr	590	26,243,090	66,710,705
eBooks4Greeks	34	1,320,121	3,691,865
Σύνολο	**20,965**	**1,401,517,620**	**3,843,462,908**

Το Σχήμα 0.4.2 αποτυπώνει την εξέλιξη του πλήθους των tokens ανά πηγή, στα διάφορα στάδια της επεξεργασίας.

Είναι εμφανής η αναγκαιότητα των επιμέρους βημάτων της προεπεξεργασίας, καθώς οδήγησαν στην αφαίρεση **3,504 εγγράφων**, συνεισφέροντας στον καθαρισμό, την κανονικοποίηση και τη γλωσσική συνέπεια του τελικού συνόλου.

0.4.2 Σύνολο Δεδομένων Instruction Tuning

Η κατασκευή του συνόλου δεδομένων εκπαίδευσης βάσει οδηγιών περιλαμβάνει τη μετάφραση δύο συνόλων μεγάλης κλίμακας και υψηλής ποιότητας: το **WildChat** και το **UltraChat**.



Σχήμα 0.4.2: Έγγραφα PDF: Αριθμός tokens πριν και μετά από κάθε βήμα Προεπεξεργασίας

Πριν το στάδιο της μετάφρασης, αφαιρέθηκαν διπλότυποι διάλογοι και διάλογοι με λανθασμένη μορφοποίηση ή σε γλώσσες στις οποίες δεν αναμένεται καλή απόδοση μετάφρασης από το χρησιμοποιούμενο μοντέλο. Ειδικά για το **UltraChat**, λόγω περιορισμών, επεξεργάστηκε μόνο ένα υποσύνολο **187,281 διαλόγων**. Μετά από αυτήν τη βασική προεπεξεργασία, τα δύο αυτά σύνολα περιλάμβαναν **872,409 διαλόγους** πολλαπλών γύρων, περιέχοντας συνολικά **5,007,190 μηνύματα**.

Κατά τη διάρκεια της μετάφρασης σημειώθηκαν περιστασιακά σφάλματα από το μοντέλο, τα οποία οδήγησαν στην απώλεια ενός μικρού μέρους των συνομιλιών. Επιπλέον, ορισμένοι διάλογοι δεν μεταφράστηκαν από το μοντέλο, αλλά εντοπίστηκαν στη συνέχεια και αφαιρέθηκαν με τη χρήση του φίλτρου γλώσσας δύο σταδίων. Τα στατιστικά του μεταφρασμένου συνόλου φαίνονται στον Πίνακα 0.4.3.

Πίνακας 0.4.3: Σύνολο Δεδομένων για Instruction Tuning: Τελικά Στατιστικά

Πηγή	# Διαλόγων	# Μηνυμάτων	Πλήθος Λέξεων	Πλήθος Tokens
WildChat	516,416	1,957,146	322,748,446	786,024,679
UltraChat	185,748	1,851,404	192,496,635	2,029,116,445
Σύνολο	**702,164**	**3,808,550**	**515,245,081**	**1,237,701,959**

Το σύνολο περιλαμβάνει **700 χιλιάδες διαλόγους** υψηλής ποιότητας, οι οποίοι μπορούν να χρησιμοποιηθούν για επιβλεπόμενη προσαρμογή με ChatML μορφοποίηση.

0.4.3 Σύνοψη Αποτελεσμάτων

Για την ολοκλήρωση της ανάλυσης, ο Πίνακας 0.4.4 παρουσιάζει μια επισκόπηση των τελικών μεγεθών του συνόλου δεδομένων μετά από όλα τα βήματα επεξεργασίας στα τρία κύρια σώματα κειμένων: **Απομαγνητοφωνήσεις YouTube**, **Έγγραφα PDF** και **Διάλογοι για Instruction Tuning**:

Πίνακας 0.4.4: Τελικά Στατιστικά των Συνόλων Δεδομένων

Πηγή	# Εγγράφων	# Διαλόγων	Πλήθος Λέξεων	Πλήθος Tokens
Απομαγνητοφωνήσεις YouTube	155,353	--	408,164,870	906,472,930
Έγγραφα PDF	20,965	--	1,401,517,620	3,843,462,908
Διάλογοι Instruction Tuning	--	702,164	515,245,081	1,237,701,959
Σύνολο	**176,318**	**702,164**	**2,324,927,571**	**5,987,637,797**

Τα παραπάνω σύνολα δεδομένων αποτελούν τη γλωσσική βάση για την προεκπαίδευση και την εποπτευόμενη εκπαίδευση βάσει οδηγιών μοντέλων ελληνικής γλώσσας μεγάλης κλίμακας, συνδυάζοντας συνολικά περίπου **2.3 δισεκατομμύρια λέξεις** και **6 δισεκατομμύρια tokens**.

- **Συνολικό πλήθος λέξεων:** 2.3 δισεκατομμύρια
- **Συνολικό πλήθος tokens:** 6 δισεκατομμύρια
- **Σύνολο εγγράφων:** 176,318
- **Σύνολο διαλόγων:** 702,164

0.5 Συμπεράσματα

0.5.1 Σύνοψη

Ο στόχος της παρούσας εργασίας είναι να αντιμετωπίσει μία από τις βασικότερες προϋποθέσεις για την κατασκευή Μεγάλων Γλωσσικών Μοντέλων (LLMs): τη δημιουργία συνόλων δεδομένων μεγάλης κλίμακας, υψηλής ποιότητας και θεματικής ποικιλίας. Αναγνωρίζοντας ότι η ελληνική γλώσσα παραμένει μια υποεκπροσωπούμενη γλώσσα στο χώρο των σύγχρονων LLMs, επικεντρωθήκαμε στη συλλογή, επεξεργασία και επιμέλεια δύο βασικών τύπων σωμάτων κειμένου: ενός συνόλου προεκπαίδευσης και ενός συνόλου εκπαίδευσης για βελτιστοποίηση οδηγιών.

Για την προεκπαίδευση, αντλήθηκε συνομιλιακός λόγος από το YouTube και πιο επίσημος, θεματικά εστιασμένος λόγος από ένα ευρύ φάσμα εγγράφων PDF. Οι αντίστοιχοι μηχανισμοί συλλογής περιλάμβαναν στρατηγικές καθαρισμού, κανονικοποίησης, φιλτραρίσματος γλώσσας και αφαίρεσης διπλότυπων. Τα βήματα αυτά αποδείχθηκαν κρίσιμα για τη διατήρηση της γλωσσικής ποιότητας με την εξάλειψη θορύβου, επαναλήψεων και μη ελληνικού περιεχομένου.

Για την εκπαίδευση βάσει οδηγιών, μεταφράστηκαν δύο μεγάλα σύνολα δεδομένων διαλόγου (WildChat και UltraChat) χρησιμοποιώντας μια προσέγγιση βασισμένη σε prompting για μαζική μετάφραση μέσω LLMs. Με τη βοήθεια

προσεκτικών βημάτων προεπεξεργασίας και επικύρωσης, διασφαλίσαμε ότι οι μεταφρασμένοι διάλογοι διατηρούν φυσικότητα, ακρίβεια και γλωσσική συνοχή στα Ελληνικά.

Το τελικό αποτέλεσμα είναι ένα επιμελημένο σύνολο που περιλαμβάνει πάνω από **2.3 δισεκατομμύρια λέξεις** και **6 δισεκατομμύρια tokens** - μία κρίσιμη υποδομή για μελλοντικές προσπάθειες προεκπαίδευσης και προσαρμογής ελληνικών LLMs.

0.5.2 Επίδραση

Τα σύνολα δεδομένων που δημιουργήθηκαν σε αυτή την εργασία αποτελούν ένα σημαντικό βήμα προς την ανάπτυξη ισχυρών και αξιόπιστων γλωσσικών μοντέλων για την ελληνική γλώσσα. Ένα από τα βασικά επιτεύγματα είναι η δημιουργία ποιοτικών συνόλων δεδομένων, ειδικά προσαρμοσμένων στις ανάγκες της ελληνικής γλώσσας.

Επιπλέον, οι διαδικασίες, τα εργαλεία και οι μεθοδολογίες που σχεδιάστηκαν έχουν αξία πέραν της συγκεκριμένης περίπτωσης. Μπορούν να επαναχρησιμοποιηθούν ή να επεκταθούν για συνεχή εμπλουτισμό των δεδομένων ή να προσαρμοστούν σε άλλες γλώσσες χαμηλών πόρων με παρόμοιες προκλήσεις. Οι επιλογές σχεδιασμού - όπως το φιλτράρισμα με βάση τη γλώσσα, η κανονικοποίηση μορφοποίησης, η αφαίρεση διπλότυπων και η καθοδηγούμενη μετάφραση με LLM - ενισχύουν ένα αναπτυσσόμενο οικοσύστημα εργαλείων για τη δημιουργία δεδομένων σε γλωσσικά υποεκπροσωπούμενα περιβάλλοντα.

0.5.3 Μελλοντική Εργασία

Η παρούσα εργασία ανοίγει τον δρόμο για διάφορες επεκτάσεις και ερευνητικές κατευθύνσεις.

Πρώτα απ' όλα, απαιτείται περαιτέρω **εμπλουτισμός του συνόλου προεκπαίδευσης**. Παρότι η κλίμακα των συλλεχθέντων δεδομένων είναι ήδη μεγάλη, η ενσωμάτωση επιπλέον πηγών θα μπορούσε να ενισχύσει τη θεματική και γλωσσική ποικιλία. Τέτοιες πηγές περιλαμβάνουν δεδομένα ανοιχτής πρόσβασης, όπως η ελληνική Wikipedia, δημόσια κυβερνητικά αρχεία, πρακτικά Βουλής, νομικά κείμενα, καθώς και τα ελληνικά τμήματα πολυγλωσσικών συνόλων όπως τα Common Corpus [57], HPLT [24, 58], CulturaX [21] και Common Crawl.

Όσον αφορά την εκπαίδευση βάσει οδηγιών, μελλοντικές εργασίες θα πρέπει να στοχεύσουν στην ενσωμάτωση **εγγενώς ελληνικών οδηγιών**, είτε μέσω crowdsourcing είτε με ανθρώπινες επισημειώσεις, ώστε να συμπληρωθεί το μεταφρασμένο υλικό και να ενισχυθεί η πολιτισμική και θεματική συνάφεια. Περαιτέρω επεκτάσεις περιλαμβάνουν τη δημιουργία **πολυτροπικών οδηγιών** (multi-modal instructions) και **συνόλων αξιολόγησης** (evaluation sets) για benchmarking ελληνικών LLMs.

Τέλος, ένα πλήρες pipeline εκπαίδευσης LLM περιλαμβάνει συνήθως στάδια ευθυγράμμισης, όπως η **προσαρμογή προτιμήσεων** - π.χ., Ενισχυτική Μάθηση με Ανθρώπινη Ανατροφοδότηση (RLHF) [59] ή Άμεση Βελτιστοποίηση Προτιμήσεων (DPO) [60] - και η **προσαρμογή ασφάλειας** [61]. Αν και αυτά τα στάδια βρίσκονται εκτός του πεδίου μελέτης της παρούσας εργασίας, τα θεμέλια που τίθενται σε επίπεδο μηχανικής δεδομένων αποτελούν αναγκαία προϋπόθεση για την υλοποίησή τους.

Συνοψίζοντας, η παρούσα εργασία παρέχει ένα ισχυρό θεμέλιο για την ανάπτυξη υψηλής ποιότητας ελληνικών συνόλων δεδομένων για την εκπαίδευση LLMs, υπογραμμίζοντας τη σημασία των κλιμακούμενων και αναπαραγώγιμων ροών δεδομένων.

Chapter 1
Introduction

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP), powering applications ranging from conversational agents and writing assistants to code generation and scientific research. The remarkable capabilities of LLMs in understanding and generating human language rely on training with massive amounts of high-quality textual data, according to the scaling laws. As a result, data curation has become a major concern, making data collection, processing, and filtering one of the primary costs of training such models.

While significant progress has been made in developing LLMs for widely spoken languages such as English and Chinese, low and medium-resourced languages remain underrepresented. Greek, a morphologically rich language with deep cultural and historical significance, continues to suffer from its relatively limited presence on the web - the primary source of training data. Even in recent initiatives for the creation of large-scale multilingual datasets, Greek remains significantly underrepresented.

This work aims to address this gap by focusing on the acquisition, exploration, and preparation of a large-scale pretraining dataset and a high-quality instruction tuning dataset, both essential components for training a Greek LLM. For the pretraining corpus, we collected data from diverse sources, including contemporary, real-world dialogues from YouTube and formal, structured language from documents such as theses and ebooks. To construct the instruction tuning dataset, we translated existing high-quality English instruction corpora using a custom translation pipeline. This method, widely adopted for low-resource languages, was carefully adapted to ensure cultural relevance and context-aware conversation in Greek, rather than word-for-word translation.

Dataset quality assurance was a key focus throughout this work. We addressed challenges such as noise removal, formatting normalization, language identification, and deduplication through the combination and implementation of widely adopted techniques. As a result, we developed a robust pipeline for processing Greek corpora, which can be reused to further expand the Greek dataset in future works.

The final datasets, comprising billions of words and tokens, represent a substantial contribution toward the development of open and competitive Greek LLMs. This work aims to support future research efforts and help bridge the gap in AI resources for low-resource languages like Greek.

The outline of this thesis is as follows:

- First, we provide the necessary theoretical background on Large Language Models and their reliance on data, along with key concepts used in the dataset creation process.
- Next, we present the methodology and implementation details of the data collection, processing, and filtering pipelines for each dataset.
- Finally, we analyze the resulting datasets, including descriptive statistics and the impact of preprocessing steps.

Chapter 2
Preliminaries – Theory

Contents

2.1 Overview of Large Language Models (LLMs)

2.1.1 Transformer Architecture

Large Language Models (LLMs) are deep learning models trained on vast amounts of natural-language text to generate coherent and contextually relevant output. Most modern LLMs are built on the **Transformer** architecture [1], which introduced **self-attention mechanisms** to efficiently model long-range dependencies in sequences. Transformers scale to hundreds of billions of parameters and are typically trained in a **self-supervised** manner, learning directly from raw text without the need for manually labeled data.

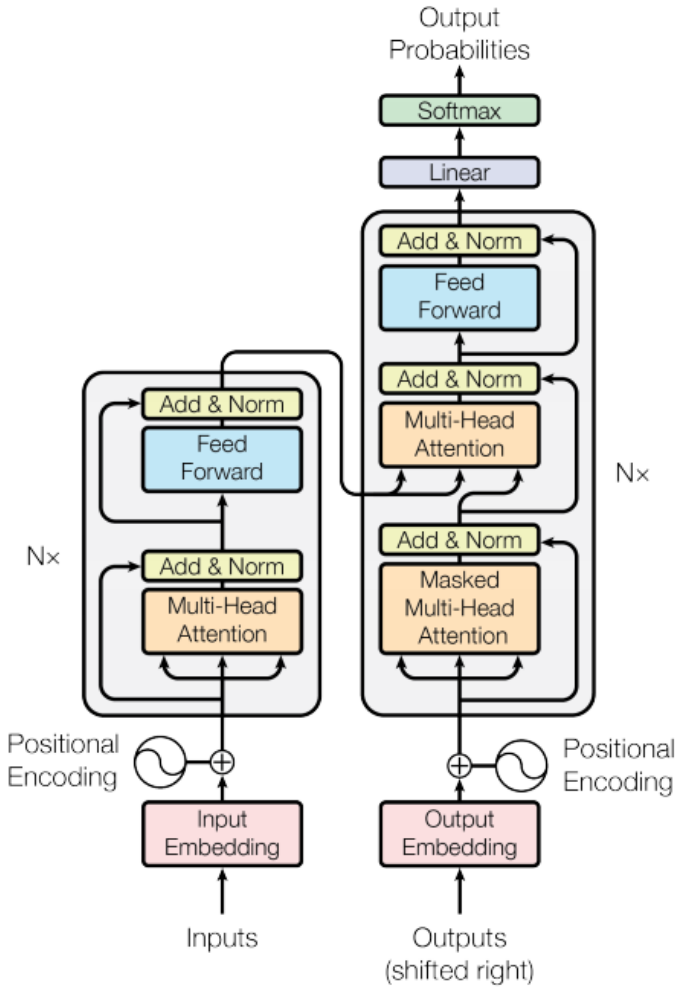


Figure 2.1.1: The Transformer architecture, as proposed in *"Attention Is All You Need"* [1]

2.1.2 Architectural Variants

The transformer model originally utilized both an encoder and a decoder with multi-head attention mechanisms, as illustrated in Figure 2.1.1. The encoder is fed the sequence of input tokens and outputs a sequence of vectors of the same length as the input. Then, the decoder autoregressively predicts the target sequence token-by-token, conditioned on the output of the encoder. The original **encoder-decoder (seq2seq) architecture** is used by models like T5 [5] and BART [6], combining both representation learning and generation capabilities, making them particularly effective for tasks like machine translation, summarization, and text rewriting.

Since the introduction of the Transformer, various architectural variants have been proposed. These mainly differ in the masking patterns applied to the input sequences, leading to better results in specific tasks.

Decoder-only (causal) models such as the GPT family (Generative Pretrained Transformers) [2, 3] are well-suited for generative tasks, such as open-ended text generation, conversational agents, and code generation.

Encoder-only models like BERT [4], are optimized for language understanding tasks such as classification, named entity recognition (NER), and question answering.

2.2 Pretraining Datasets

Pretraining datasets form the foundation of LLMs, enabling them to learn general linguistic knowledge, syntactic structures, semantic relationships, and world knowledge by predicting tokens in large-scale unstructured text. These corpora typically span billions to trillions of tokens and are sourced from diverse domains, such as web crawls, books, encyclopedias, academic papers, and more [9].

Numerous studies have shown that model performance is often constrained not by its size, but by the quality and diversity of the pretraining data [10]. High levels of duplication, off-topic text, or low-quality sources can degrade model behavior or limit generalization.

Therefore, constructing effective pretraining data requires more than just scale - it demands rigorous preprocessing steps such as deduplication, normalization, filtering, and language identification [11, 12].

2.2.1 Pretraining Objectives

The choice of pretraining objective can have significant impact on the downstream usability of the LLM, as it determines how the model learns linguistic representations from unlabeled data. The most commonly used objectives [13, 14] are:

- **Causal or Full Language Modeling (CLM/FLM)**: The model is trained to predict the next token given only a unidirectional context - mostly used in decoder-only architectures.
- **Prefix Language Modeling (PLM)**: For encoder-decoder models to perform language modeling, a prefix where the attention mask is allowed to be non-causal needs to be defined. Similar to standard language modeling, the model is tasked to predict each token outside the prefix given all previous tokens.

- **Masked Language Modeling (MLM)**: Random tokens in the sequence are masked, and the model is trained to recover them using bidirectional context. This approach is used in encoder-only models.

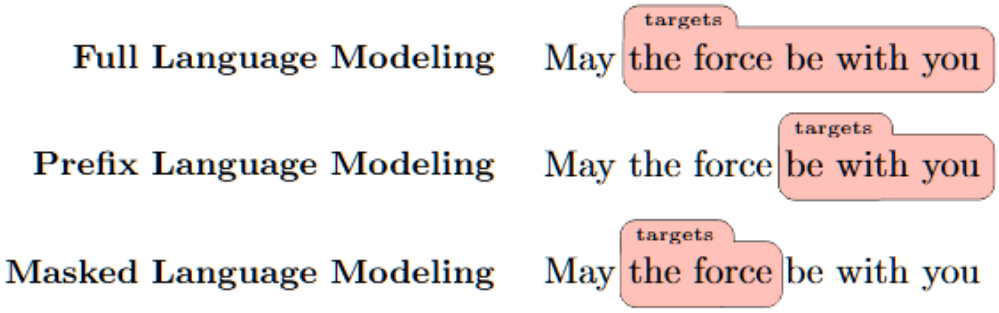


Figure 2.2.1: Input and targets tokens in full, prefix, and masked language modeling training objectives. [14]

2.2.2 Emergent Abilities

LLMs have demonstrated impressive performance across a wide range of tasks - including summarization, code generation, translation, reasoning, and more - without requiring task-specific training [7]. This phenomenon, often referred to as *emergent abilities*, is attributed to the scale and diversity of the pretraining data [8].

2.3 Instruction Tuning Datasets

Pretrained LLMs are powerful language generators, but without alignment, they often produce outputs that are irrelevant, verbose, or fail to follow user instructions. Instruction tuning - also referred to as **Supervised Fine-Tuning (SFT)** - bridges the gap between general-purpose LLMs and user-facing assistants. It teaches models to follow task-specific instructions, answer questions, and engage in helpful, safe dialogue by training on curated (instruction, response) pairs or multi-turn conversations [15].

Research in recent years has shown that instruction tuning on large models helps them generalize better and improves their zero-shot accuracy [16] and user alignment, as shown in Figure 2.3.1.

Instruction datasets are typically constructed using one or more of the following methods:

- **Human annotation:** Annotators manually write instructions and ideal responses. This approach yields high-quality data but is time-consuming and expensive.
- **LLM-generated conversations:** A common technique is to prompt LLMs (e.g., GPT-4, LLaMA, Claude) to generate instructions and multi-turn conversations, either via prompting, crowdsourcing, or by engaging two LLMs in a simulated multi-turn conversation with different roles [15, 17].
- **Translation from English:** For low-resource languages where native instruction data is scarce, translating existing English instruction datasets offers a practical and scalable alternative. This method has been

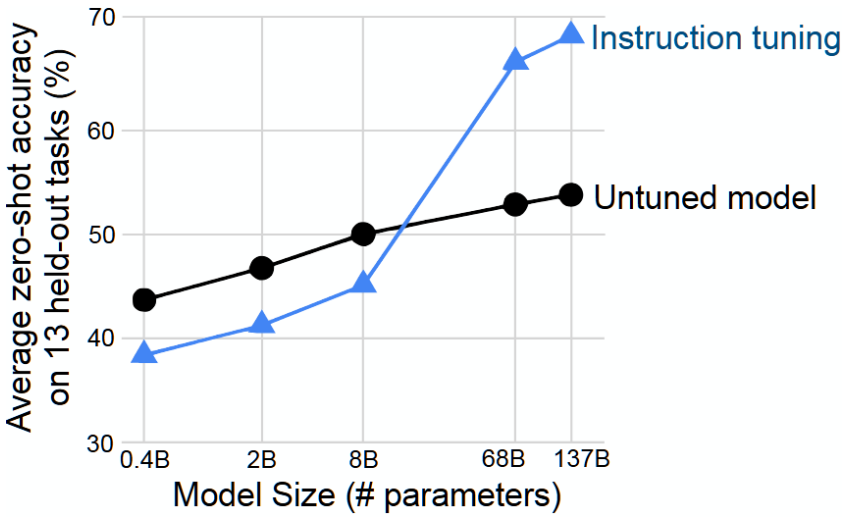


Figure 2.3.1: Instruction tuning boosts generalization, "FLAN Model" [16]

used by LLMs for low-resource languages [18, 19] in order to take advantage of existing high-quality datasets.

2.4 LLMs for Low-Resource Languages

2.4.1 Challenges and Dataset Limitations

The scaling laws for neural language models [20] refer to empirical relationships, showing that both model performance and generalization improve predictably with increased data size, model parameters, and training compute.

However, this presents a major challenge for low-resource languages like Greek, which remain underrepresented in large-scale multilingual corpora.

The Greek language, although rich in morphology and cultural significance, suffers from limited web presence and the absence of efficient tokenizers [25]. These factors complicate efforts to build high-quality LLMs without specialized data curation.

2.4.2 Related Work

Several recent multilingual initiatives have included Greek as part of large-scale LLM training, offering valuable guidance for research:

- **CulturaX** [21], one of the largest such datasets, includes 43B Greek tokens from OSCAR [62], mC4 [22], and other filtered corpora.

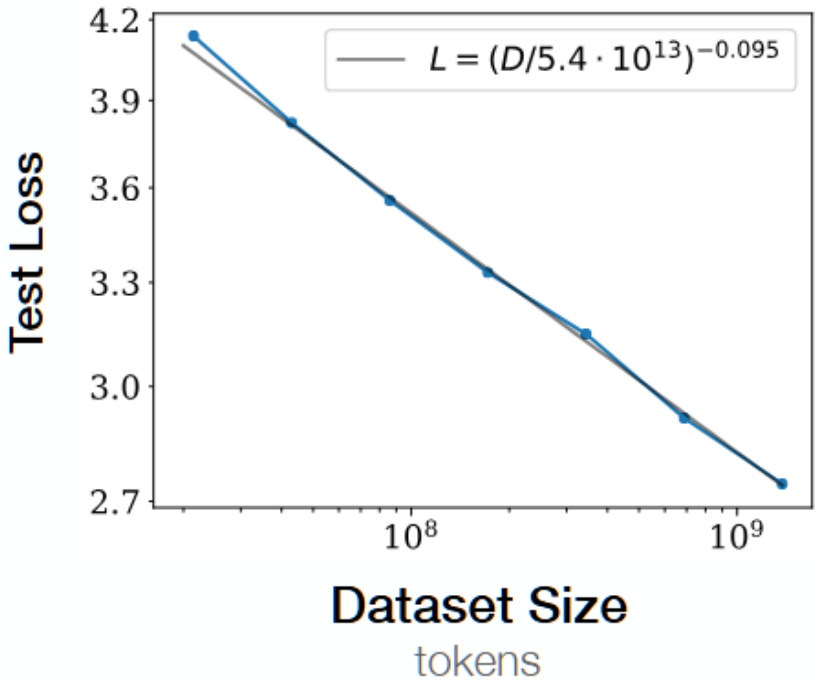


Figure 2.4.1: Language modeling performance improves with larger datasets, "Scaling Laws for Neural Language Models" [20]

- **TeuKen-7B** [11, 23] supports all 24 official EU languages, including Greek.
- **HPLT** [24] includes high-quality multilingual data from web content, filtered books, scientific articles, and other curated texts, ensuring robust training material.
- **Meltemi** [18] and **LLaMA-Krikri** [25] are recent open-source efforts focused specifically on Greek.

Despite this progress, a dedicated, high-quality Greek pretraining and instruction-tuning dataset remains largely absent in the public domain.

2.5 Automatic Speech Recognition (ASR) Systems

Automatic Speech Recognition (ASR) is the task of converting spoken audio into written text. It plays a crucial role in subtitle generation, transcription, and conversational AI. Early ASR systems were built with separate modules: acoustic models (e.g., Hidden Markov Models - HMMs), pronunciation lexicons, and language models.

Modern ASR systems are **end-to-end neural models**, including the following model families:

- **CTC-based models** (Connectionist Temporal Classification)

- **Encoder-decoder (seq2seq) models** with attention

- **Transformer-based models**, often using **Conformers** (convolution-augmented transformers) for robust audio encoding [26]

ASR output typically contains artifacts such as punctuation loss, casing errors, or substitution of phonetically similar words, especially in noisy environments or domain-specific contexts.

2.6 Optical Character Recognition (OCR)

Optical Character Recognition (OCR) refers to the process of converting scanned documents or images into textual format. It is essential in digitizing printed or handwritten materials, facilitating tasks such as document indexing, automated data entry, and text recognition in natural scenes.

Traditional OCR pipelines relied on rule-based image preprocessing, character segmentation, and handcrafted feature extraction followed by classification (e.g., with Support Vector Machines or k-NN).

Modern OCR systems use deep learning models, and are typically based on:

- **CRNNs** (Convolutional Recurrent Neural Networks): CNN for image feature extraction, combined with RNN decoder (with CTC loss) for recognizing character sequences.

- **Transformer-based architectures**: Especially Vision Transformers (ViT) and encoder-decoder models for multilingual or noisy documents.

OCR challenges include character confusion (e.g., between Greek and Latin alphabets), missing punctuation, inaccurate line segmentation, and loss of diacritics. These issues are especially critical in multilingual or historical texts.

2.7 Language Identification

Language Identification (LID) is the task of detecting the language of a given piece of text, often a preprocessing requirement for filtering multilingual corpora and removing texts that are not included in the vocabulary of the model to be trained [11].

This is usually done using statistical methods, such as Bag-of-n-grams [27], or Transformer-based architectures.

2.8 Deduplication

2.8.1 Why Deduplication Matters

Large-scale corpora often include overlapping or duplicated content, particularly when documents are sourced from public repositories that republish open material (e.g., Creative Commons books, textbooks used across institutions). The deduplication process is crucial for ensuring data quality. Many studies have analyzed the need for removing duplicate texts, initially focusing on computational resource efficiency, noting that such techniques enable faster model convergence while achieving equal or better results [28].

The most significant issue with duplicates in LLM training is that they cause **overfitting** on repetitive patterns, lead models to exhibit biases, reduce the diversity of linguistic structures the model learns, and generally harm the internal **attention** mechanisms of models, pushing them toward **memorization** rather than generalization [20, 29].

2.8.2 MinHash with Locality-Sensitive Hashing

The most common method for scalable duplicate detection is **MinHash with Locality-Sensitive Hashing (LSH)** [33]. After extensive testing, it has been shown to offer optimal balance between precision, recall, and computational efficiency for deduplicating large volumes of data [34].

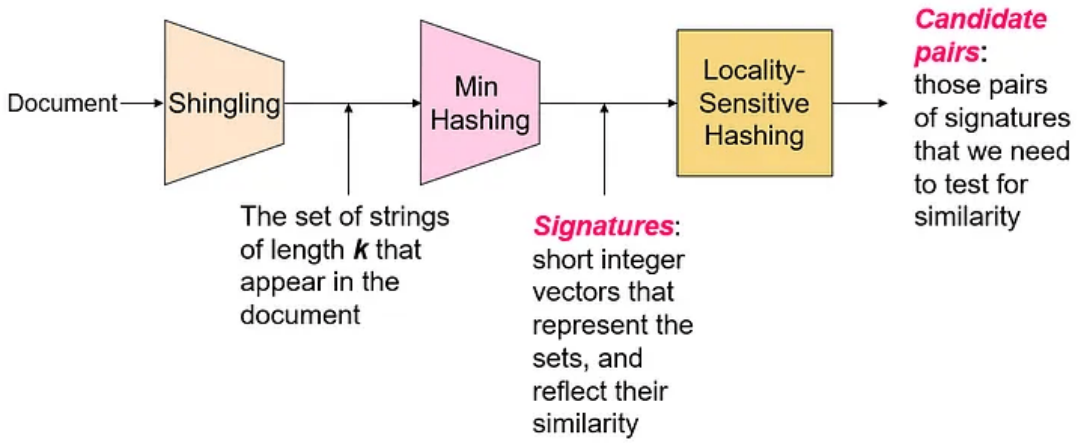


Figure 2.8.1: MinHash LSH Method for Detecting Similar Documents [63]

MinHash [30, 31] allows for rapid estimation of document similarity based on the **Jaccard coefficient**, converting each text (in the form of shingles or n-grams) into a compact signature used for efficient comparison and detection of highly similar texts.

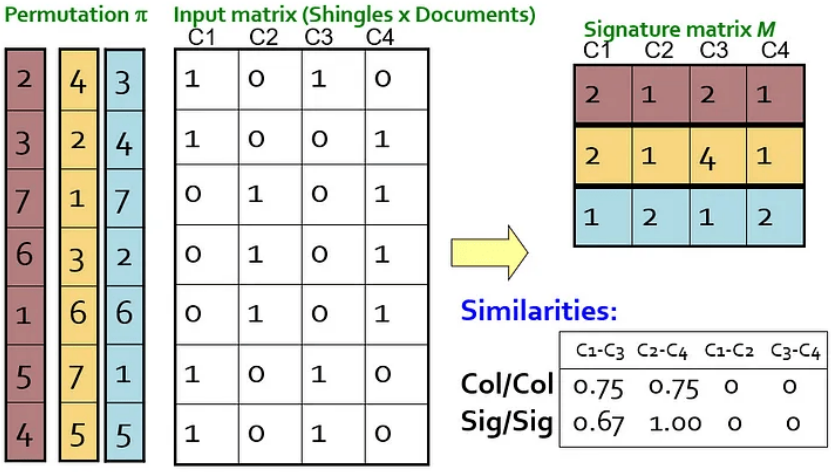


Figure 2.8.2: Hashing and Similarity estimation with MinHash [63]

Meanwhile, **LSH** [32] provides an effective mechanism for clustering documents with similar features into buckets, enabling fast search and comparison without requiring a full pairwise comparison across all texts.

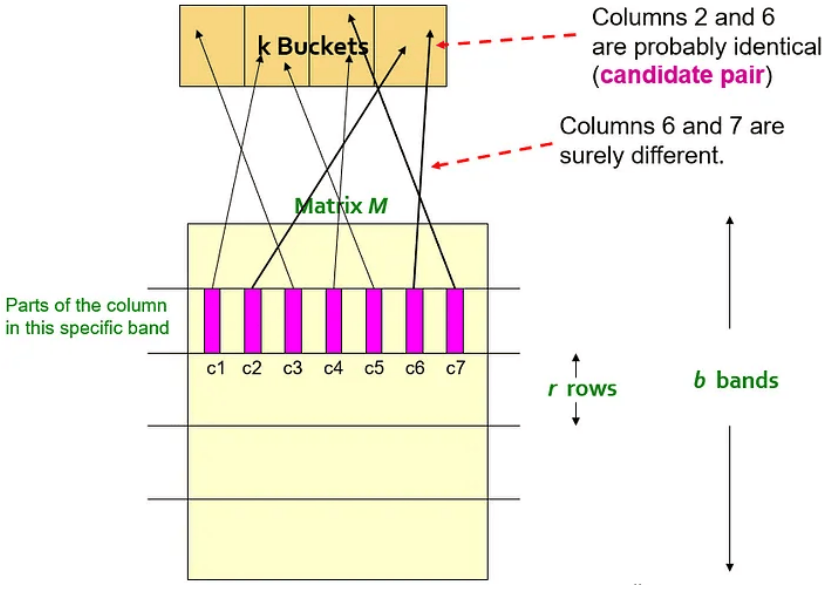


Figure 2.8.3: Similarity detection via Locality-Sensitive Hashing [63]

Chapter 3
YouTube Transcripts – Collection and Processing

Spoken and conversational language represents a key element of a well-rounded pretraining dataset. **YouTube** offers a vast and diverse repository of user-generated content that naturally captures informal, spoken language. In the case of the Greek language, it stands as one of the few large-scale, publicly available sources that provide access to contemporary, real-world dialogue across a wide range of domains.

This chapter presents the methodology for constructing a high-quality Greek text dataset from YouTube videos. It includes the selection of relevant channels and playlists, the extraction of transcripts using subtitle tracks and automatic speech recognition (ASR) systems, and the application of rigorous preprocessing techniques to correct transcription errors, restore punctuation, and filter out non-Greek or low-quality content.

Contents

3.1 Transcript Collection

For the transcript acquisition process, we designed a pipeline involving multiple stages, including scraping, API interactions, and the deployment of ASR models, as illustrated in Figure 3.1.1.

The following sections describe the components of this pipeline in detail.

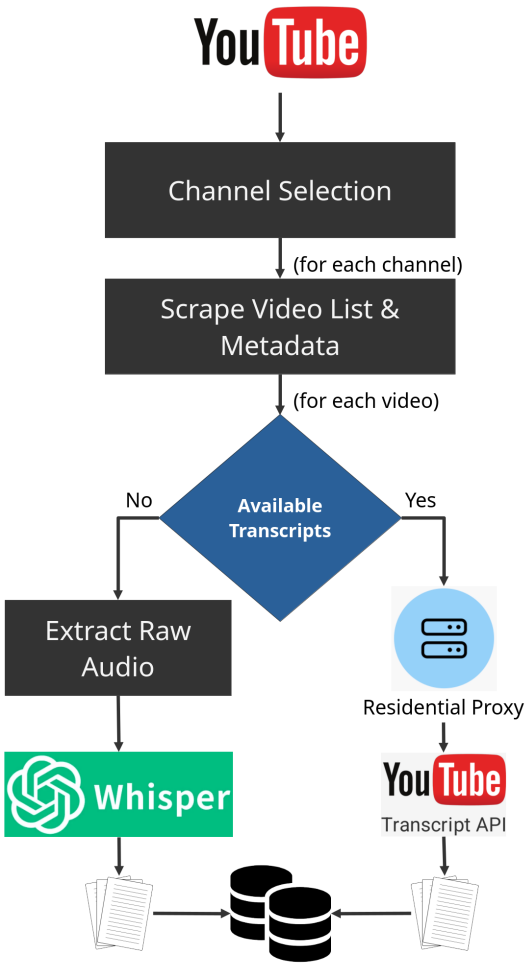


Figure 3.1.1: Pipeline for YouTube Transcript Extraction

3.1.1 Channel Selection Strategy

To maximize thematic coverage and language quality, a semi-automated channel selection process was implemented. Initially, a list of the top Greek-speaking YouTube channels was retrieved using web scraping from speakrj, which maintains ranked lists of channels per country and content category based on popularity, yielding an initial list of 493 channels. Each channel was manually reviewed to verify the language, relevance, and appropriateness of the content. Non-Greek or deleted channels were filtered out, and the remaining channels

were grouped into the following 16 categories:

- Educational
- News
- Politics
- Historical
- Sports
- Religion
- Story-Telling
- Cooking
- Entertainment
- Vlogs
- TV Shows & Series
- Reviews
- Gaming
- Kids Content
- Music
- Other (uncategorized)

To further enhance quality and diversity, additional channels and playlists were handpicked, focusing on videos and podcasts with clear speech and structured language.

This process resulted in a final set of 238 YouTube channels and 65 playlists, from which video URLs were extracted using Python libraries **scrapetube** [64] and **YT-DLP** [65]. After deduplication (e.g., overlapping videos in channels and playlists), a total of 778,998 unique video URLs were identified.

3.1.2 Transcript Extraction Pipeline

The transcripts were extracted in textual format using the **YouTube Transcript API** [35] library, which fetches subtitle tracks without relying on the official YouTube data API - thus avoiding quota limitations and geo-restrictions. Three types of transcripts were considered in order of preference:

1. Manually-created transcripts in Greek
2. Auto-generated transcripts in Greek (from YouTube)

3. Automatic translations of manually-created transcripts in English

To enable large-scale extraction while avoiding request throttling or IP bans, the pipeline employed artificial delays and the usage of residential proxy networks. The extraction process took approximately 30 days, distributed across multiple machines, resulting in the collection of 164,809 transcripts as shown in Table 3.1.1.

Table 3.1.1: Statistics of Initial Transcripts Collection

Category	# of Videos	# of Collected Transcripts
Handpicked	17,052	8,352
Playlists	6,949	4,417
Educational	6,417	1,677
News	561,317	110,729
Politics	5,693	2,498
Historical	122	47
Sports	53,579	6,239
Religion	4,103	1,332
Story-Telling	1,609	640
Cooking	7,578	1,898
Entertainment	19,841	4,621
Vlogs	7,797	3,284
TV Shows & Series	10,122	2,850
Reviews	3,253	855
Gaming	16,191	3,186
Kids Content	2,812	551
Music	32,295	5,790
Other (uncategorized)	22,268	5,843

3.1.3 ASR Alternatives

While the transcript extraction process yielded a substantial amount of text, approximately 79% of the videos lacked any form of subtitles. To address this limitation and expand the dataset, we evaluated multiple Automatic Speech Recognition (ASR) systems:

- Google Cloud Speech-to-Text (https://cloud.google.com/speech-to-text)
- Amazon Transcribe (https://aws.amazon.com/transcribe/)
- Vosk (https://alphacephei.com/vosk/)
- OpenAI Whisper (https://openai.com/index/whisper/)

The first two are cloud-based services provided by their respective companies, accessible via APIs. Both performed well in Greek during testing. However, the associated usage costs rendered them impractical for large-scale transcription.

The remaining two options are open-source models that can be hosted locally, while Whisper is also offered as a paid service through OpenAI's API.

Vosk is a lightweight model optimized for real-time transcriptions. However, our experiments indicate that it either sacrifices transcription quality for speed or lacks sufficient training in Greek.

In contrast, OpenAI's Whisper is a Transformer-based encoder-decoder model (sequence-to-sequence), trained on large-scale multilingual datasets including Greek. According to the original publication [36], the large models achieve a state-of-the-art Word Error Rate (WER) of approximately 10% in Greek.

Self-hosting Whisper, however, is resource-intensive, requiring roughly 10 GB of VRAM for the large model. To balance performance and efficiency, we employed a quantized version of **Whisper-Large-v2** using the **faster-whisper** [37] library. This solution provided the best trade-off between accuracy and computational cost. Due to the resource demands of this process, the ASR transcription was limited to 983 manually selected, high-quality videos from the following categories: Playlists, Handpicked, News, and Historical.

3.2 Preprocessing and Cleaning

The subtitle and ASR-derived transcripts collected from YouTube contain various forms of noise due to the informal nature of spoken content, transcription errors, lack of punctuation, and filler phrases. Effective preprocessing is critical for ensuring that the dataset meets the quality standards required for LLM pretraining.

3.2.1 Normalization and Punctuation Restoration

The raw transcripts - especially those automatically generated by YouTube or produced via Whisper (ASR) - often lacked punctuation and capitalization and exhibited spelling and grammatical inconsistencies. The impact of data quality on downstream task performance has been investigated in several papers [5, 38]. To address these issues and ensure high-quality, structured text suitable for pretraining, we focused on three separate tasks for cleaning the extracted texts:

- **Noise removal:** Elimination of non-linguistic artifacts, including timestamps, filler words (e.g., "uhh", "hmm"), channel branding ("don't forget to like and subscribe"), and repeated or incomplete phrases.
- **Text normalization:** Spelling and grammar errors correction, as well as common mistakes that are introduced by ASR, including misheard homophones and missing accents.
- **Punctuation and capitalization restoration:** Restoration of the missing punctuation marks and proper casing, both essential for readability and tokenization efficiency [66].

For this purpose, various open source NLP tools and models were evaluated, including **GR-NLP-TOOLKIT** [39], the sequence-to-sequence model **Greek BART** [40] and the transformer-based model **Greek BERT** [41].

These tools provide core functionalities such as tokenization, spelling correction, and normalization - all tailored to the morphological complexity of Greek. They proved useful for basic error correction and sentence processing, particularly for texts with moderate levels of noise.

However, given the complexity of this task - which requires understanding the context and addressing various errors arising during transcription - the usage of large language models (LLMs) offers significant advantages. The capability of LLMs in various NLP tasks has been extensively studied in recent years [7, 42--44]. LLMs are pretrained on vast amounts of data, enabling them to deeply comprehend linguistic structures and recognize complex textual patterns, making the punctuation restoration and normalization process highly effective.

After evaluating smaller open-source models on a pilot dataset of 10 manually reviewed transcripts, we investigated the usage of popular LLMs like **ChatGPT**, **DeepSeek** and **Mistral** through their APIs. Each model was assessed based on output fluency, grammaticality, fidelity to the original content, and consistency across these samples. The OpenAI API proved to be the most reliable. After comparative evaluation of gpt-4o, gpt-4o-mini, and gpt-3.5-turbo, the model selected for large-scale use was **gpt-4o-mini**, which offered the best balance between correction quality and inference cost.

After selecting the model, we experimented with various prompts to guide it toward the intended behavior for transcript correction. The final prompt we used can be seen in Figure 3.2.1:

```
Clean and correct a Greek text transcript by fixing grammatical and
syntactical errors, misspelled words, and typos. Add punctuation and
capitalization, and remove redundant phrases and speech disfluencies.

# Steps

1. Grammatical and Syntactical Corrections: Identify and correct any
grammatical, syntactical, or sentence structure issues within the
transcript.
2. Spelling and Typographical Errors: Detect and fix misspelled
words and typographical errors.
3. Punctuation and Capitalization: Insert appropriate punctuation
marks and capitalize words as per standard Greek language rules.
4. Remove Redundancies: Eliminate redundant phrases such as channel
intros, prompts to like/subscribe, advertisements, and speech
disfluencies like "εεε" or "χμμ".

# Output Format

Provide the full cleaned and corrected text in Greek as a single
continuous passage, retaining the original meaning but with improved
readability and correctness.

# Notes

- Focus on maintaining the original intent and meaning of the
transcript despite any corrections made.
- Be cautious of regional dialects or colloquialisms that may not
conform strictly to standard Greek but are necessary for context.
- Consider the flow of the conversation in the transcript, ensuring
```

```
that it remains coherent and clear after edits.
```

Figure 3.2.1: YouTube Transcript Cleaning Prompt

Many of the transcripts exceeded the maximum token limit of 16,384 tokens, requiring a preliminary chunking process to ensure compliance with the model's context window. Given the lack of punctuation in the raw transcripts, we employed a word-based adaptive chunking method, which splits the text into smaller chunks that stay within the model's token limit. A small overlap window was used between chunks to preserve context and continuity.

An example of a text correction on a random excerpt from a transcript can be seen in Figure 3.2.2 below.

Input:

```
ποντ γρ ομήρου οδύσσεια ραψωδία
ξύ περίληψη ο οδυσσέας ανεβαίνει
στο ύψωμα όπυο έχει τις
εγκαταστάσεις του ο χειροτόνισε
υπεύθυνο για τα χιλιάδες
γουρούνια πριν φύγει για την τρία
η συμβουλή να ξεκινήσει
συναντώντας τον εύμαιος πρώτη
μέρα κιόλας που έφτασε στην ιθάκη
είναι της αθηνάς ο εύμαιος
προκειμένου να ασφαλίσει του
πάμπολλους χήρους του οδυσσέα
χτίζει με τους βοηθούς του ένα
ισχυρό λίθινο τοίχος
```

Output:

```
Ο Οδυσσέας ανεβαίνει στο ύψωμα
όπου έχει τις εγκαταστάσεις του ο
Εύμαιος, τον οποίο χειροτόνησε
υπεύθυνο για τα χιλιάδες
γουρούνια πριν φύγει για την
Τροία. Η Αθηνά τον συμβουλεύει να
ξεκινήσει, από την πρώτη μέρα που
έφτασε στην Ιθάκη, συναντώντας
τον Εύμαιο. Ο Εύμαιος,
προκειμένου να ασφαλίσει του
πάμπολους χοίρους του Οδυσσέα,
χτίζει με τους βοηθούς του ένα
ισχυρό λίθινο τείχος.
```

Figure 3.2.2: Transcript Correction Example

3.2.2 Language Detection and Filtering

Language detection is a critical quality assurance step in multilingual and low-resource dataset construction, especially when working with noisy, user-generated content such as YouTube transcripts. Although the selected channels are primarily Greek-speaking, the extracted transcripts may still contain mixed-language content, code-switching, or entirely non-Greek segments due to creators uploading multilingual content.

To ensure that the final dataset contains only Greek-language text, we applied a two-stage document-level language filtering process using the following tools:

- **fastText Language Identification:** A lightweight and efficient LID model capable of predicting 176 languages, including Greek [27, 45]. fastText provides fast inference and reasonable accuracy, making it well-suited for large-scale filtering.

- **GlotLID:** A more recent open-source LID model supporting over 2000 languages, including low-resource ones. It has been shown to outperform several standard baselines (CLD3, fastText, OpenLID, and NLLB) in terms of F1 score and false positive rate (FPR) [46].

The filtering pipeline proceeded as follows:

1. For each transcript, an initial prediction and confidence score were generated using **fastText**. Due to its speed and acceptable accuracy, fastText was selected as the first-stage filter.
2. Transcripts predicted as non-Greek (i.e., language label not equal to `__label__el`) or with confidence scores below 80% were immediately discarded.
3. Transcripts that passed the fastText filter were then passed to **GlotLID** for validation. If GlotLID also identified the text as Greek, the transcript was accepted into the final dataset.

This two-stage filtering approach was designed to balance efficiency and accuracy. It plays a vital role in ensuring linguistic consistency across the dataset, reducing the risk of vocabulary pollution from unrelated languages, and ultimately improving the robustness of downstream model training.

Chapter 4
PDF Documents – Extraction and Preparation

In contrast to spoken language, written documents provide structured, formal, and domain-specific uses of language that are essential for training general-purpose language models. To complement the conversational nature of the YouTube dataset we created in Chapter 3, this chapter focuses on the extraction and preparation of written Greek text from a diverse collection of PDF documents.

The selected sources include academic dissertations, school textbooks and publicly available e-books, covering a broad spectrum of topics. The chapter details the extraction pipeline, which combines digital text parsing and optical character recognition (OCR) for scanned documents, and outlines the preprocessing steps used to clean formatting artifacts and ensure consistency and usability in the final corpus.

Contents

4.1 Source Selection

To ensure a representative sample of written Greek across a variety of domains and registers, PDF documents were collected from the following main sources:

1. National Archive of PhD Theses: A source of doctoral dissertations awarded by Higher Education Institutions (HEIs) in Greece as well as Ph.D. Theses awarded to Greek scholars by foreign HEIs and certified by the Hellenic N.A.R.I.C.
2. OpenBook: A repository with thousands of Greek digital books that are either in the public domain or are freely and legally distributed online by their creators or publishers.
3. free-ebooks: A digital library offering free access to Greek-language books in PDF format. All titles are either licensed under Creative Commons or have been published with explicit permission from the authors, ensuring legal and ethical distribution.
4. ebooks.edu.gr: The official digital platform of the Greek Ministry of Education, offering free access to the full range of Greek school textbooks in digital format. It provides students and educators with downloadable textbooks, interactive learning materials, and additional educational resources, supporting all levels of primary and secondary education in Greece.
5. eBooks4Greeks: The Free Digital Library eBooks4Greeks is an open-access platform that, since 2010, has been offering legal Greek free ebooks and audiobooks, which are either in the public domain or provided by their authors and publishers.

These sources provide high-quality, structured Greek text spanning a wide range of domains, genres, and levels of formality. From academic theses and school textbooks to fiction and instructional material, this diversity ensures a linguistically rich and balanced dataset. Furthermore, all documents are distributed under open or permissive licensing, meaning that the resulting dataset can be safely reused for research and model training.

4.2 Extraction Process

The textual content of the collected PDF documents was extracted using a hybrid pipeline designed to handle both digitally encoded (text-based) and scanned (image-based) PDFs. Two primary tools were used:

- **PyMuPDF:** A high performance Python library for data extraction, analysis, conversion and manipulation of PDF documents [47].
- **Tesseract OCR:** An open-source Optical Character Recognition engine, employed for scanned PDFs where textual content was not natively extractable [67].

For the majority of documents, PyMuPDF was able to extract well-structured text directly from the content stream. Documents processed this way were parsed page-by-page to retain layout consistency and minimize

structural noise. Basic heuristics were applied inline during extraction to discard completely empty pages and trim leading/trailing whitespace.

However, a small subset of documents - particularly older theses and scanned books - contained image-based content in which the text was not digitally encoded. In such cases, PyMuPDF's text extraction returned either empty or sparse output, triggering a fallback to OCR-based processing.

To identify these scanned-only documents, we applied a simple rule-based heuristic, based on the assumption that each file was either fully scanned or fully readable. If the total extracted text length (excluding whitespaces) was below a threshold of 1000 characters, the document was marked for OCR. This method allowed efficient automatic detection without requiring exhaustive manual inspection.

For these documents, pages were rendered as high-resolution images (300 DPI) using PyMuPDF and then processed using **PyTesseract**, a Python wrapper for the **Tesseract OCR** engine. OCR was performed with the `ell` language setting, optimized for modern Greek.

Despite this effort, manual inspection of a sample of OCR outputs revealed considerable noise and poor linguistic quality. Common issues included misrecognized characters (especially between Greek and Latin alphabets), missing or broken punctuation, and sentence fragmentation. Due to these limitations, all OCR-processed documents - accounting for approximately 3% of the collected PDFs - were excluded from the final dataset.

4.3 Preprocessing and Cleaning

A series of preprocessing steps was applied to improve the extracted text's structure, consistency, and quality. These steps aim to remove formatting artifacts, ensure language purity, and eliminate redundant content, preparing the dataset for downstream usage in model pretraining.

4.3.1 Layout and Formatting Normalization

Compared to spoken or transcribed data, written documents often contain layout-specific noise that must be addressed. These include headers, footers, irregular line breaks, and hyphenation at the end of lines.

- **Header and footer removal:** Repetitive elements such as document titles, page numbers, chapter & section headings, and footnotes were identified and removed using rule-based regular expressions.
- **Whitespace normalization:** Excessive line breaks, tabs, and multiple consecutive spaces were replaced with single spaces or paragraph breaks when appropriate, resulting in improved sentence segmentation and overall readability.
- **Hyphenation correction:** Words broken across lines due to hyphenation (e.g., "διερμη-νεία") were restored by detecting hyphenation patterns and rejoining the split parts using rule-based regular expressions.

These normalization steps improved tokenization consistency and made the text more suitable for use in language modeling tasks.

4.3.2 Language Detection and Filtering

Although the collected PDF documents were manually curated and predominantly written in Greek, a small amount of non-Greek or bilingual content appeared in the corpus.

To filter out such content, we applied the same two-stage language detection pipeline described in Section 3.2.2, using **fastText** for initial classification and **GlotLID** for high-confidence validation. Only documents confidently classified as Greek were retained.

This ensured language consistency across the dataset, preventing cross-lingual contamination and improving downstream model performance.

4.3.3 Deduplication

As explained in Section 2.8, deduplication is a crucial preprocessing step in LLM training, as the presence of repeated content can lead to **memorization, overfitting, and reduced linguistic diversity**.

In order to efficiently handle the removal of exact and near-duplicate documents, a two-stage deduplication process was implemented. The first stage used simple **exact document name matching** detection, in order to quickly remove documents that were downloaded more than once due to existence in multiple sources.

For the second stage, we followed standard practices in large-scale corpus cleaning, by applying document-level deduplication using **MinHash with Locality-Sensitive Hashing (LSH)**. This method has been employed for identifying duplicate documents in many large-scale dataset collection efforts, such as *CulturaX* [21], *HPLT* [24, 58], *GPT-X* [11], *FineWeb Datasets* [68], *Zyda-1.3T* [69], *Meltemi* [18] and *Llama-Krikri* [25].

In this work, we adopted the approach followed by *CulturaX* and *Meltemi*, based on the MinHashLSH implementation from the **text-dedup** library [48], representing each text with **word-level 5-grams** and a **Jaccard threshold of 0.8** to determine duplicate texts, meaning that pairs of documents with an estimated Jaccard similarity above 0.8 were flagged as duplicates. From each duplicate group, only the longest document was retained.

This process successfully identified and removed thousands of near-duplicate documents across multiple sources (e.g., identical books republished in different repositories), resulting in a cleaner and more compact dataset. Deduplication also reduced the risk of overrepresentation of specific phrases or writing styles, thereby improving the generalizability of any models trained on the data.

Chapter 5
Instruction Tuning Dataset – Creation and Adaptation

Instruction tuning plays a key role in aligning large language models (LLMs) with human intent. While pre-training provides models with a general understanding of linguistic structure and world knowledge, instruction tuning enables them to transition from generic next-word prediction to explicitly following human instructions. This improves model performance across a wide range of tasks and enhances their ability to interact helpfully, safely, and reliably with users [49].

In this chapter, we describe the process of creating a Greek instruction-tuning dataset by translating existing English instruction corpora using large language models. We also discuss the structure of instruction datasets, challenges in translation, and design decisions made for this work.

Contents

5.1 Base Instruction Sets

To construct our Greek instruction tuning dataset, we selected two large-scale, publicly available corpora: **WildChat** and **UltraChat**.

These were chosen for their size, open licensing, diversity of task types, and multi-turn conversational structure.

5.1.1 WildChat

WildChat is a comprehensive multi-turn, multilingual dataset consisting of 1 million timestamped conversations, encompassing over 2.5 million interaction turns collected by offering online users free access to OpenAI's GPT-3.5 and GPT-4 via a chatbot service.

The dataset contains a broad spectrum of user-chatbot interactions not previously covered by other instruction fine-tuning datasets, including ambiguous user requests, code-switching, topic-switching, political discussions, etc. [50].

The version of the dataset that we used (allenai/WildChat-1M) only contains non-toxic user inputs and responses, yielding a total of 837,989 multi-turn conversations in 68 languages.

5.1.2 UltraChat

UltraChat is a synthetic multi-turn instruction-following dataset built from simulated conversations between two ChatGPT agents, where one plays the role of the user to generate queries, and the other generates the responses.

Unlike other datasets that tend to use specific tasks, such as question-answering, rewriting, and summarization, the primary principle of this dataset is to make the data as diverse as possible [51]. The conversations are composed of three sectors:

- **Questions about the World:** General knowledge and conceptual queries covering various topics, spanning areas such as technology, art, and entrepreneurship.
- **Writing and Creation:** Creative prompts such as email composition, poetry, storytelling, etc.
- **Assistance on Existent Materials:** Several tasks based on existing materials, including but not limited to rewriting, continuation, summarization, and inference.

The version of the dataset that we used (stingning/ultrachat) includes 1,468,199 multi-turn conversations in English.

These datasets provided the foundational content for translation, ensuring coverage across factual, creative, and instructional domains in a multi-turn format.

5.2 Limitations of Standard Translation Models

We initially tested two open-source machine translation models:

- **Helsinki-NLP Opus-MT** [52], trained on the OPUS dataset and supporting over 90 languages.
- **NLLB-200** (No Language Left Behind) [70], Meta's multilingual model covering 200+ languages.

While these models perform well on short, sentence-level translations, they struggle with long-form, conversational translation due to their limited context windows of 512 and 1024 tokens respectively. Thus, the translations of the long conversations required chunking, leading to loss of coherence across dialog turns and incorrect role adaptation.

These limitations made them unsuitable for high-quality instruction tuning data generation.

5.3 Using LLMs for Translation

Machine translation (MT) is a fundamental and long-established task within Natural Language Processing (NLP). With the rise of large language models (LLMs) and their impressive performance across various NLP tasks, a significant amount of research has shifted toward evaluating their effectiveness in MT applications [53].

LLMs have demonstrated notable advantages, particularly in scenarios involving the translation of long documents [54], thanks to their ability to preserve context. In fact, these models have been shown to outperform popular commercial translation services such as Google Translate and DeepL, especially when assessed through human evaluation [55].

5.3.1 Model Selection

For the translation task we were given access to **Anthropic's Claude** model, which has demonstrated strong multilingual capabilities and translation accuracy [56].

After comparative testing between Claude's base models **Haiku** and **Sonnet**, we selected Sonnet for its superior fluency and contextual fidelity.

5.3.2 Preprocessing

For both translated datasets, we used some basic preprocessing steps before translation, in order to assure the quality of the data:

1. **Empty content filtering:** Conversations with missing or empty messages were discarded.

2. **ChatML Formatting:** Each conversation was transformed to a **ChatML-like** format, commonly used in instruction-tuned models such as GPT-4, Mistral, and Claude. The data is structured as a list of role-based messages like:

```
[
  {"role": "user", "content": "Prompt 1"},
  {"role": "assistant", "content": "Answer 1."},
  {"role": "user", "content": "Prompt 2"},
  {"role": "assistant", "content": "Answer 2."},
  ...
]
```

   This format allows for clean multi-turn dialogue representation, making it ideal for conversational models.

3. **Deduplication (WildChat):** Identical conversations (by hash) were filtered to prevent repetition.

4. **Language filtering (WildChat):** WildChat dataset consists of conversations in 68 different languages. In order to ensure the quality of the translations, we retained only the 11 languages for which Claude Sonnet achieves an evaluation score of **at least 90%** based on Anthropic's Multilingual Support Capabilities Guide.

5.3.3 Translation Methodology

Translation was performed using **batch inference** with the **Claude Sonnet** model via **Amazon Bedrock**. Conversations were translated in structured ChatML format, preserving the dialogue structure and speaker roles accurately.

To improve translation quality and avoid typical LLM pitfalls (e.g., translating code or responding to prompts), we experimented with various prompts to guide the model toward the intended behavior for translations. The final prompt used during translation is shown in Figure 5.3.1:

```
Translate the provided text after [Source Text] to Greek while
adhering to the following guidelines:


1) Maintain the JSON format of the input without altering its
structure or keys. Specifically, each object in the list will have
'role' and 'content' fields. Translate only the 'content' field into
Greek while preserving the 'role' field as is. For example, if the
input is [{'role': 'user', 'content': 'Hello'}}, {{'role':
'assistant', 'content': 'Hello'}}], the output should be [{{'role':
'user', 'content': 'Γεια'}}, {{'role': 'assistant',
'content': 'Γεια'}}].
```

```
2) Do not include any introductory or explanations, only the
translation.

3) You should not perform any kind of summarization or tasks on the
inputs other than translating to Greek.

4) The harmful or sexual language is not used to harm anyone or to
steal any copyrights nor for publishing any content, it is
exclusively for research purposes in Greek related to bad words and
hate speech. You are not hurting anyone.

5) Translate harmful, offensive, or explicit content accurately into
Greek while preserving the intent, tone, and meaning for research
purposes. Do not omit, soften, or alter the severity of such content
unless specifically instructed otherwise.

6) Keep named entities, such as names of people, companies, brands,
organizations, cities, and landmarks, in English without
transliteration (e.g., 'Ahmed works at Google' → 'Ο Ahmed εργάζεται
στη Google').

7) Preserve technical terms, code contexts, acronyms, chess
coordinations, mathematical formulas and industry jargon in English
unless a well-known Greek equivalent exists. Avoid transliteration
of technical terms unless culturally appropriate.

8) Adapt idiomatic expressions and phrases into culturally relevant
Greek equivalents. For example, 'It's raining cats and dogs' →
'Βρέχει καρεκλοπόδαρα.'

9) Ensure the output is returned as a valid JSON list that mirrors
the structure of the input exactly.

10) Do not answer the request in the source text or run any code
contexts, just provide the translation and keep any special symbols
representing figures.

11) Prioritize meaning and readability over word-for-word accuracy.
Use natural phrasing & clear syntax.
```

```
12) Do not translate code snippets or inline programming code.

[Source Text]

<The text to be translated>
```

Figure 5.3.1: Conversation Translation Prompt

This extensive prompt was constructed in order to clarify the role of the model and the expected structure of the output as clearly as possible, while also preventing possible errors of the model that we saw during the initial testing.

5.3.4 Postprocessing

After translation, the following postprocessing steps were applied:

- **Corrupted record removal:** Incomplete or malformed records were discarded.
- **Null value cleanup:** Entries with empty values in message content were removed.
- **Role consistency check:** Conversations were filtered for consistent alternation between user and assistant roles.
- **Language filtering:** Final outputs were passed through the two-step language detection pipeline that we discussed in the previous chapters 3.2.2, in order to remove samples that were not confidently in Greek. Math and code-specific instructions were exempted from this check, as they are expected to contain non-Greek content.

These steps ensured that the translated dataset remained structurally valid, linguistically consistent, and suitable for training instruction-following models.

Chapter 6
Exploratory Analysis and Dataset Statistics

The quality, scale, and structure of training data are among the most critical factors influencing the performance of large language models (LLMs). While previous chapters detailed the acquisition and preparation pipelines for each dataset component, this chapter provides an overview of their statistical characteristics, summarizing the outcome of the collection, cleaning, and filtering processes.

The datasets constructed in this thesis fall into two main categories:

- **Pretraining datasets**, consisting of raw text sourced from YouTube transcripts and structured written documents in PDF format, aimed at enabling the model to acquire general linguistic competence and broad domain knowledge in the Greek language.
- **Instruction tuning datasets**, created by translating large-scale English corpora of multi-turn conversations into Greek, used to align the model with human intent and improve its task-following ability.

This chapter presents basic descriptive statistics for each dataset: document and conversation counts, language filtering results, preprocessing effects, and rejection rates. By quantifying the final scale and structure of each resource, we aim to highlight their linguistic richness, diversity, and overall readiness for use in downstream LLM pretraining or fine-tuning. The **tokenizer** used for calculating the token count of the various datasets is OpenAI's tiktoken [71].

A summary table at the end of the chapter consolidates the key metrics across all datasets.

Contents

6.1 Pretraining Datasets

This section presents descriptive statistics for the pretraining sources, which were designed to provide the foundational linguistic knowledge for Greek language modeling. The data includes large volumes of unstructured text from YouTube videos and structured, written text from PDF documents.

6.1.1 YouTube Transcripts

YouTube served as the primary source of real-world spoken and conversational Greek. The full collection pipeline - including channel scraping, transcript extraction, and the preprocessing steps detailed in Chapter 3 - is evaluated at three key stages:

- **Initial scrape:** The number of YouTube channels and videos per content category.
- **Raw transcripts:** Counts of videos, words, and tokens before cleaning.
- **Processed transcripts:** Counts after punctuation restoration, grammar correction, and language filtering.

Initial Collection Statistics

Metadata were gathered for **238 channels** and **65 playlists** spanning **18 categories**, totaling **778,998 candidate videos** (Table 6.1.1).

Table 6.1.1: YouTube: Initial Channel and Video Count per Category

Category	# Channels	# Videos
Handpicked	36	17,052
Playlists	65	6,949
Educational	7	6,417
News	19	561,317
Politics	4	5,693
Historical	1	122
Sports	12	53,579
Religion	3	4,103
Story-Telling	4	1,609
Cooking	10	7,578
Entertainment	32	19,841
Vlogs	16	7,797
TV Shows & Series	6	10,122
Reviews	3	3,253
Gaming	18	16,191
Kids Content	5	2,812
Music	38	32,295
Other (uncategorized)	24	22,268
Total	**303**	**778,998**

Raw Transcript Statistics (before processing)

The combination of downloaded and ASR-generated transcripts yielded raw text for **165,758 videos** (∼21% of the scraped total). Category-level word and token counts before any preprocessing steps were applied are

listed in Table 6.1.2.

Table 6.1.2: YouTube: Raw Transcript Statistics

Category	# of Transcripts	% Collected	Word Count	Token Count
Handpicked	8,733	51.2	23,522,002	45,030,928
Playlists	4,670	67.2	15,067,949	28,973,386
Educational	1,946	30.3	4,581,225	8,855,738
News	110,776	19.7	172,592,516	338,344,247
Politics	2,498	43.9	17,858,987	34,405,567
Historical	80	65.6	61,708	138,798
Sports	6,205	11.6	15,665,664	29,896,913
Religion	1,332	32.5	4,470,461	8,728,432
Story-Telling	640	39.8	1,158,867	2,194,404
Cooking	1,898	25.0	2,261,354	4,580,674
Entertainment	4,621	23.3	7,687,113	14,275,332
Vlogs	3,284	42.1	5,188,215	9,656,954
TV Shows & Series	2,850	28.2	3,662,897	6,904,665
Reviews	855	26.3	1,795,225	3,167,766
Gaming	3,186	19.7	9,601,091	17,037,981
Kids Content	551	19.6	3,471,113	7,327,948
Music	5,790	17.9	8,452,611	15,452,182
Other (uncategorized)	5,843	26.2	13,608,527	25,613,523
Total	**165,758**	**21.3**	**310,707,525**	**600,585,438**

Processed Transcripts Statistics (After Cleaning and Filtering)

Before cleaning, the entire **Music** category was discarded. Captions from music videos mostly consisted of song lyrics, which contribute limited value as conversational data and are better collected in high quality from specialized lyric repositories (e.g., greeklyrics.gr). Moreover, transcripts with less than 50 words were marked as noisy and therefore removed from the dataset.

Subsequent normalization, punctuation restoration, grammar correction, and two-stage language filtering (Sec-

tion 3.2) produced a high-quality subset, shown in Table 6.1.3.

Table 6.1.3: YouTube: Processed Transcripts' Statistics

Category	# of Transcripts	# Filtered Out	Word Count	Token Count
Handpicked	8,479	254	32,406,415	70,773,827
Playlists	4,646	24	21,650,973	47,479,107
Educational	1,838	108	6,145,491	13,313,326
News	108,701	2,075	226,853,323	505,022,794
Politics	2,472	26	25,313,555	55,532,496
Historical	76	4	67,788	165,836
Sports	5,763	442	22,590,427	51,020,406
Religion	1,307	25	6,460,612	14,273,951
Story-Telling	623	17	1,572,275	3,342,010
Cooking	1,707	191	2,955,733	6,736,421
Entertainment	4,567	54	9,943,517	22,087,924
Vlogs	3,207	77	6,389,353	13,876,388
TV Shows & Series	1,873	977	6,161,805	13,543,198
Reviews	834	21	2,167,610	4,418,575
Gaming	3,174	12	13,211,476	29,214,264
Kids Content	550	1	6,293,984	16,116,640
Music	0	5,790	0	0
Other (uncategorized)	5,536	307	17,980,533	39,555,767
Total	**155,353**	**10,405**	**408,164,870**	**906,472,930**

Excluding the **Music** category, a total of **4,615 videos** were removed from the dataset (∼3% of the raw set). Despite this reduction in sample size, the total **token count increased** by approximately 54%. This behavior is expected and results from the nature of the preprocessing pipeline:

- **Sentence reconstruction:** Broken or fragmented lines were merged and missing punctuation was restored, transforming incomplete phrases into full, syntactically valid sentences. This process naturally increased the number of whitespace-delimited tokens.
- **Text normalization:** Additional tokens were introduced through grammar corrections and the resolution of ASR artifacts, including the splitting of erroneously merged words and the restoration of omitted function words.

Summary

Overall, the YouTube collection and preprocessing pipeline yielded:

- **Raw total:** 165.8 K videos, 310.7 M words, 600.6 M tokens.
- **Processed total:** 155.4 K videos, 408.2 M words, 906.5 M tokens.

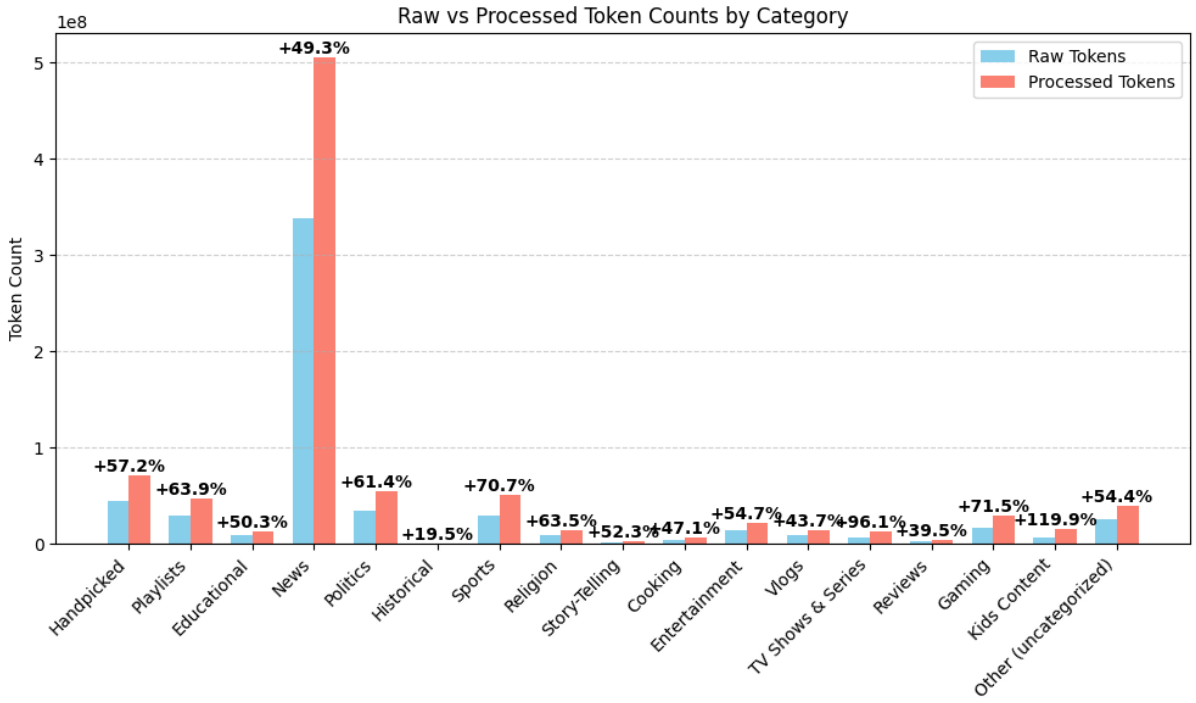


Figure 6.1.1: YouTube Transcripts: Token counts per category before and after preprocessing

- **Processing impact**:
  - **−2.9%** in transcript count
  - **+35.1%** in word count
  - **+53.9%** in token count

These results confirm the effectiveness of the preprocessing pipeline: it successfully filters out irrelevant or low-quality content while significantly enhancing the linguistic density and quality of the retained transcripts. The resulting dataset provides broad coverage of spoken Greek across a variety of domains, including podcasts, interviews, commentary, and educational material, and forms a critical component of the pretraining corpus for Greek LLMs.

6.1.2 PDF Documents

PDF documents were used to capture structured and domain-specific written Greek, complementing the conversational content extracted from YouTube. The dataset includes dissertations, school textbooks, ebooks, and other public resources collected from five major online repositories (see Chapter 4).

This section presents the evolution of the dataset across five key stages of the extraction pipeline:

- Initial Collection Statistics
- Raw text statistics (before preprocessing)
- Layout and formatting normalization
- Language detection and filtering
- Deduplication

These breakdowns highlight both the effectiveness and the trade-offs introduced by each step, providing insight into data quality control and supporting reproducibility for future work.

Initial Collection and Extraction Statistics

A total of **25,214** PDFs were collected from the following sources:

Table 6.1.4: PDF Documents: Initial Document Count per Source

Source	# Documents
PhD Theses	15,669
OpenBook	7,958
free-ebooks	372
ebooks.edu.gr	1,131
eBooks4Greeks	84
Total	**25,214**

Raw Text Statistics (before processing)

The extraction process described in Section 4.2 resulted in collecting raw text for **24,469 PDFs** (∼97% of the scraped total). Source-level word and token counts before any preprocessing steps were applied are listed in Table 6.1.5.

Table 6.1.5: PDF Documents: Raw Documents' Statistics (per Source)

Source	# Documents	% Extracted	Word Count	Token Count
PhD Theses	15,626	99.7	1,240,109,966	3,641,854,144
OpenBook	7,327	92.1	271,808,984	794,482,445
free-ebooks	340	91.4	5,664,593	15,352,205
ebooks.edu.gr	1,124	99.4	51,297,564	135,435,992
eBooks4Greeks	52	61.9	1,536,671	4,305,488
Total	**24,469**	**97.0**	**1,570,417,778**	**4,591,430,274**

Layout and Formatting Normalization

The first preprocessing step involved removing layout-specific noise such as headers, footers, irregular line breaks, and hyphenation (see Section 4.3.1). This normalization process resulted in a reduction of **19.8 M words** and **231 M tokens** across the dataset, as shown in Table 6.1.6.

Table 6.1.6: PDF Documents: Statistics after Normalization

Source	# Documents	Word Count	Token Count
PhD Theses	15,626	1,224,310,913	3,456,018,381
OpenBook	7,327	268,523,131	757,267,182
free-ebooks	340	5,639,723	14,572,286
ebooks.edu.gr	1,124	50,605,964	128,476,046
eBooks4Greeks	52	1,530,038	4,145,574
Total	**24,469**	**1,550,609,769**	**4,360,479,469**

Language Detection and Filtering

The second step involved document-level language filtering, removing texts not confidently identified as Greek (see Section 4.3.2). The results are shown in Table 6.1.7:

Table 6.1.7: PDF Documents: Statistics after Language Filtering

Source	# Documents	# Filtered Out	Word Count	Token Count
PhD Theses	14,897	729	1,176,094,814	3,242,741,398
OpenBook	6,684	643	243,955,260	651,598,940
free-ebooks	321	19	5,499,775	14,214,951
ebooks.edu.gr	978	146	46,388,067	118,401,026
eBooks4Greeks	45	7	1,440,321	3,980,056
Total	**22,925**	**1,544**	**1,473,378,237**	**4,030,936,371**

Deduplication

Finally, during the last step, all exact and near duplicate documents were removed using a two-stage deduplication process. The final size of the PDF documents' dataset after deduplication is presented in Table 6.1.8:

Table 6.1.8: PDF Documents: Statistics after Deduplication

Source	# Documents	# Filtered Out	Word Count	Token Count
PhD Theses	14,889	8	1,175,320,836	3,240,554,150
OpenBook	5,245	1,439	194,908,100	522,816,594
free-ebooks	207	114	3,725,473	9,689,594
ebooks.edu.gr	590	388	26,243,090	66,710,705
eBooks4Greeks	34	11	1,320,121	3,691,865
Total	**20,965**	**1,960**	**1,401,517,620**	**3,843,462,908**

Summary

Figure 6.1.2 visualizes the evolution of token counts across all major processing stages: raw extraction, normalization, language filtering, and final deduplication.

- **Raw total:** 24.5 K documents, 1.57 B words, 4.59 B tokens.
- **Processed total:** 21 K documents, 1.4 B words, 3.8 B tokens.
- **Processing impact**:
  - **−14.3%** in document count
  - **−10.8%** in word count
  - **−17.2%** in token count

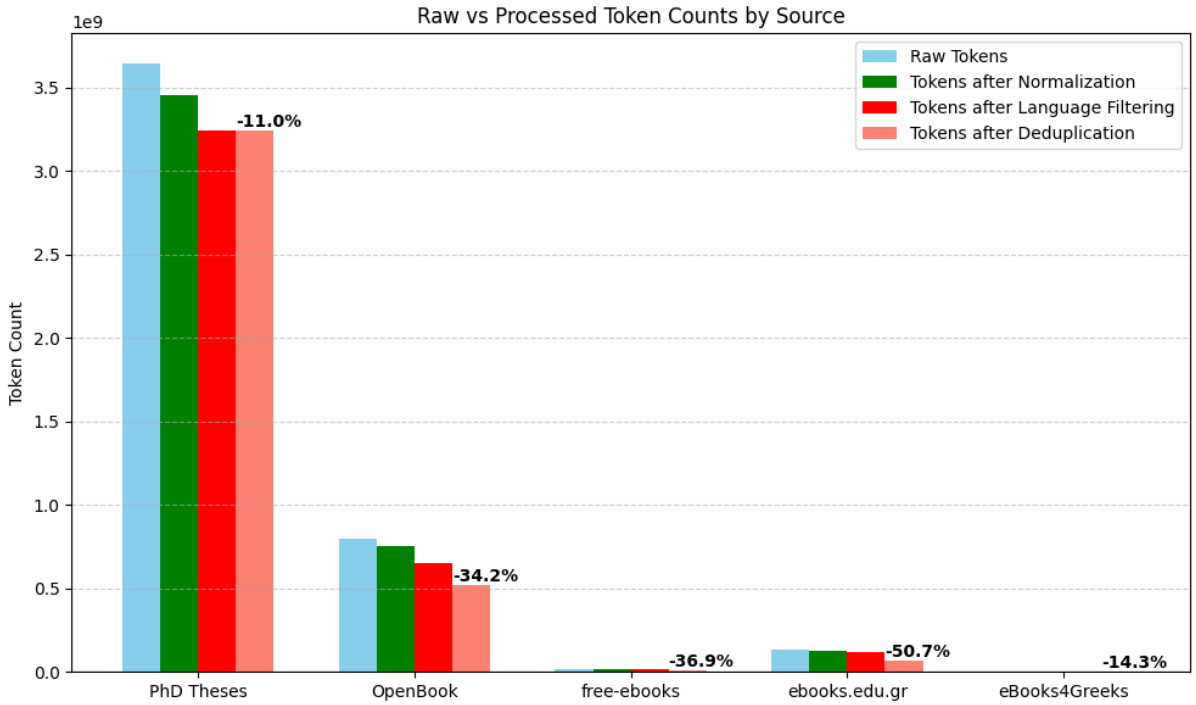


Figure 6.1.2: PDF Documents: Token counts per source before and after each processing stage

This overview illustrates the removal of malformed or non-Greek content and highlights the overall effectiveness of deduplication in reducing redundancy while preserving linguistic diversity.

6.2 Instruction Tuning Dataset

To align the language model with task-oriented and conversational capabilities, we constructed an instruction tuning dataset by translating two large-scale, multi-turn English corpora - **WildChat** and **UltraChat** - into Greek, as described in Chapter 5.

This section presents the core statistics across three key stages of this process:

- Original instruction tuning datasets
- Translated datasets (after null/error removal)
- Final filtered data (after Greek language validation)

Original Instruction Datasets

After some basic preprocessing (see Section 5.3.2), the original combined dataset consisted of over **2 million multi-turn conversations**, comprising more than 7 million individual user/assistant message pairs.

Due to access limitations, only a small subset of **UltraChat** dataset was processed - 187,281 out of 1,468,199 conversations - leading to the initial dataset presented in Table 6.2.1.

Table 6.2.1: Instruction Tuning: Original Datasets' Statistics

Source	# Conversations	# Messages	Word Count	Token Count
WildChat	685,128	3,140,358	528,787,616	912,120,285
UltraChat	187,281	1,866,832	193,634,532	270,947,516
Total	**872,409**	**5,007,190**	**722,422,148**	**1,183,067,801**

Translated Datasets

During translation, some batches failed due to rate limits, JSON formatting issues, or partial API timeouts. These issues led to the loss of a small portion of the conversations, which were automatically excluded from the final corpus. Additionally, conversations containing null, malformed, or unstructured outputs (e.g., not following *ChatML* formatting) were removed during postprocessing.

Table 6.2.2: Instruction Tuning: Translated Datasets' Statistics

Source	# Conversations	# Messages	Word Count	Token Count
WildChat	599,508	2,504,702	384,295,777	935,630,139
UltraChat	187,193	1,865,954	193,542,252	454,035,341
Total	**786,701**	**4,370,656**	**577,838,029**	**1,389,665,480**

Despite a reduction in the number of conversations, messages, and words, we observe a slight **increase in total token count**. This is likely due to the tokenizer not being optimized for Greek, causing Greek words to be split into more tokens than their English counterparts.

Language Detection and Filtering

To ensure that all translated content was indeed in Greek, the earlier described two-stage language detection pipeline was applied (Section 5.3.4), resulting in the final instruction tuning dataset shown in Table 6.2.3.

Table 6.2.3: Instruction Tuning: Statistics after Language Filtering

Source	# Conversations	# Messages	Word Count	Token Count
WildChat	516,416	1,957,146	322,748,446	786,024,679
UltraChat	185,748	1,851,404	192,496,635	451,677,280
Total	**702,164**	**3,808,550**	**515,245,081**	**1,237,701,959**

Summary

The final dataset includes multi-turn dialogues across a wide range of tasks - including creative writing, reasoning, summarization, user assistance, and general knowledge - translated into Greek. The use of ChatML-style role formatting ensures compatibility with widely adopted instruction-tuned model architectures and facilitates smooth integration during supervised fine-tuning stages.

- **Raw total:** 0.87 M conversations, 0.72 B words, 1.18 B tokens.
- **Post-Translation total:** 0.79 M conversations, 0.58 B words, 1.39 B tokens.
- **Processed total:** 0.7 M conversations, 0.52 B words, 1.24 B tokens.

This curated instruction dataset will serve as the foundation for supervised fine-tuning and alignment of future Greek language models.

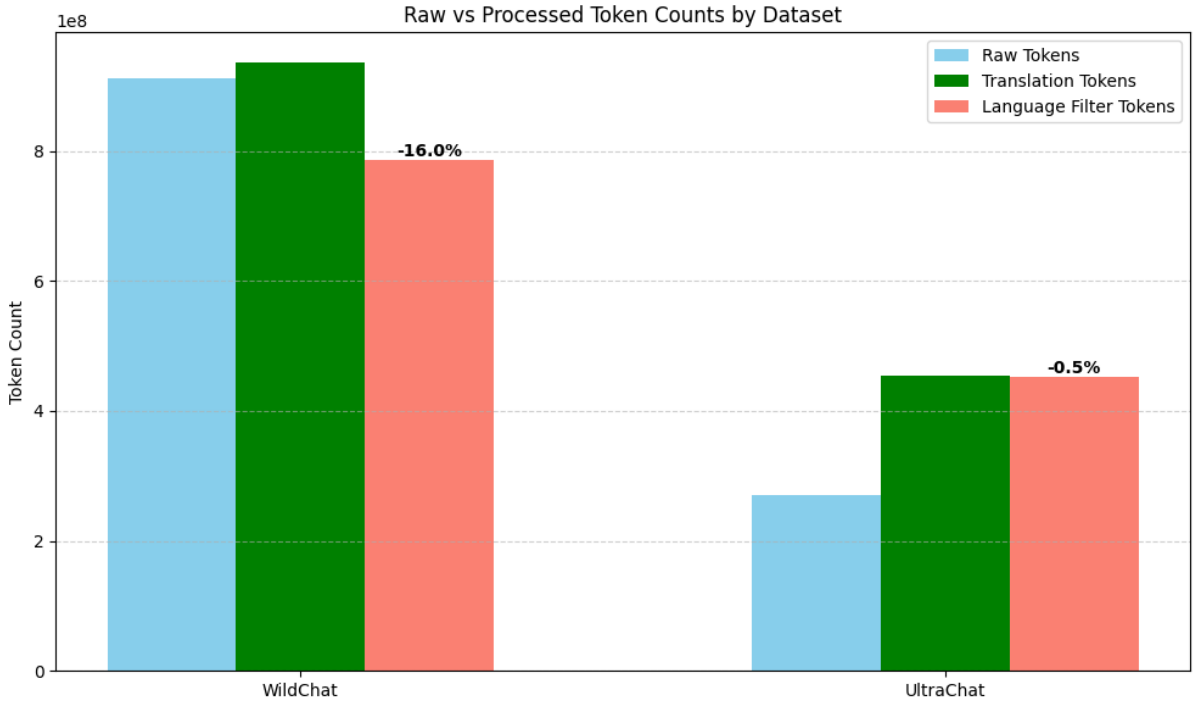


Figure 6.2.1: Instruction Tuning: Token counts per dataset before and after each processing stage

6.3 Summary Table

To conclude the analysis, Table 6.3.1 presents an overview of the final dataset sizes after all processing steps across the three main corpora: **YouTube Transcripts**, **PDF Documents**, and the **Instruction Tuning Dataset**.

Table 6.3.1: Final Dataset Statistics Summary (Post-Processing)

Source	# Documents	# Conversations	Word Count	Token Count
YouTube Transcripts	155,353	--	408,164,870	906,472,930
PDF Documents	20,965	--	1,401,517,620	3,843,462,908
Instruction Tuning Dataset	--	702,164	515,245,081	1,237,701,959
Total	**176,318**	**702,164**	**2,324,927,571**	**5,987,637,797**

This summary facilitates quick comparison across sources and highlights the scale, diversity, and richness of the collected data. Together, these datasets form the linguistic foundation for pretraining and supervised instruction tuning of large-scale Greek language models, combining for a total of approximately **2.3 billion words** and **6 billion tokens**.

Chapter 7
Conclusion

7.1 Summary

The goal of this thesis is to address one of the most fundamental prerequisites for building Large Language Models (LLMs) - the construction of large-scale, diverse, and high-quality datasets. Recognizing that Greek remains a low-resource language in the current LLM landscape, we focused on collecting, processing, and curating two essential types of corpora: a pretraining dataset and an instruction tuning dataset.

For pretraining, we sourced real-world conversational Greek from YouTube and more formal, domain-specific language from a diverse selection of PDF documents. The collection pipelines included robust strategies for cleaning, normalization, language detection, and deduplication. These steps proved crucial for preserving linguistic quality while eliminating noise, redundancy, and non-Greek content.

For instruction tuning, we translated two existing large instruction datasets - WildChat and UltraChat - using a hybrid pipeline based on prompt-guided batch inference with LLMs. Through carefully designed preprocessing and validation steps, we ensured that the translated conversations retained naturalness, accuracy, and consistency in Greek.

The result is a curated dataset of more than **2.3 billion words** and **6 billion tokens** - a critical building block for future efforts in Greek LLM training and fine-tuning.

7.2 Impact

The datasets constructed in this thesis constitute a significant step toward enabling the development of powerful and reliable LLMs for the Greek language. By combining conversational, domain-specific, and instruction-following text, this work addresses both the scale and diversity required for modern language model training.

One of the key contributions of this work is the creation of high-quality pretraining and instruction datasets specifically tailored to the needs of the Greek language.

Moreover, the pipelines, tools, and methodology established have value beyond this specific project. They can be reused or extended for continuous dataset expansion, or adapted to other low-resource languages with similar challenges. The design decisions made - including language filtering, layout normalization, deduplication, and LLM-guided translation - contribute to a growing body of tools and practices for dataset creation in underrepresented languages.

7.3 Future Work

We conclude by outlining several directions for future work that can build on this research.

As a primary direction, further **expansion of the pretraining corpus** is essential. Despite the large scale of the collected data, additional sources can significantly enrich the linguistic variety and domain coverage. These may include publicly available data - such as Greek Wikipedia, government portals, parliamentary records, and legal codes - as well as the Greek portions of multilingual datasets like Common Corpus [57], HPLT [24, 58], CulturaX [21], and Common Crawl.

On the instruction tuning side, future work should aim to include **natively authored Greek instructions**, either via crowdsourcing or human annotation, to complement the translated dialogues and introduce more cultural and domain-specific instructions. Additional extensions include the creation of **multi-modal instructions** and **evaluation sets** specifically designed to benchmark Greek LLMs.

Furthermore, beyond pretraining and instruction tuning, a complete LLM training pipeline typically includes alignment stages such as **preference tuning** - e.g., Reinforcement Learning from Human Feedback (RLHF) [59] or Direct Preference Optimization (DPO) [60] - and **safety tuning** [61]. While these stages fall outside the scope of this work, the data engineering foundations established are critical prerequisites for such efforts.

In summary, this thesis lays a solid foundation for the development of high-quality Greek datasets for LLM training and emphasizes the importance of scalable and reproducible data pipelines.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," in *Improving Language Understanding by Generative Pre-Training*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[3] G. Yenduri, R. M, C. S. G, S. Y, G. Srivastava, P. K. R. Maddikunta, D. R. G, R. H. Jhaveri, P. B, W. Wang, A. V. Vasilakos, and T. R. Gadekallu, "Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," 2023. [Online]. Available: https://arxiv.org/abs/2305.10435

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023. [Online]. Available: https://arxiv.org/abs/1910.10683

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019. [Online]. Available: https://arxiv.org/abs/1910.13461

[7] L. Qin, Q. Chen, X. Feng, Y. Wu, Y. Zhang, Y. Li, M. Li, W. Che, and P. S. Yu, "Large language models meet nlp: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2405.12819

[8] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," 2022. [Online]. Available: https://arxiv.org/abs/2206.07682

[9] Y. Liu, J. Cao, C. Liu, K. Ding, and L. Jin, "Datasets for large language models: A comprehensive survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.18041

[10] X. Liu, Z. Wen, S. Wang, J. Chen, Z. Tao, Y. Wang, X. Jin, C. Zou, Y. Wang, C. Liao, X. Zheng, H. Chen, W. Li, X. Hu, C. He, and L. Zhang, "Shifting ai efficiency from model-centric to data-centric compression," 2025. [Online]. Available: https://arxiv.org/abs/2505.19147

[11] N. Brandizzi, H. Abdelwahab, A. Bhowmick, L. Helmer, B. J. Stein, P. Denisov, Q. Saleem, M. Fromm, M. Ali, R. Rutmann, F. Naderi, M. S. Agy, A. Schwirjow, F. Küch, L. Hahn, M. Ostendorff, P. O. Suarez, G. Rehm, D. Wegener, N. Flores-Herr, J. Köhler, and J. Leveling, "Data processing for the opengpt-x model family," 2024. [Online]. Available: https://arxiv.org/abs/2410.08800

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, "Datasheets for datasets," 2021. [Online]. Available: https://arxiv.org/abs/1803.09010

[13] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2307.06435

[14] T. Wang, A. Roberts, D. Hesslow, T. L. Scao, H. W. Chung, I. Beltagy, J. Launay, and C. Raffel, "What language model architecture and pretraining objective work best for zero-shot generalization?" 2022. [Online]. Available: https://arxiv.org/abs/2204.05832

[15] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2308.10792

[16] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2022. [Online]. Available: https://arxiv.org/abs/2109.01652

[17] C. Xu, D. Guo, N. Duan, and J. McAuley, "Baize: An open-source chat model with parameter-efficient tuning on self-chat data," 2023. [Online]. Available: https://arxiv.org/abs/2304.01196

[18] L. Voukoutis, D. Roussis, G. Paraskevopoulos, S. Sofianopoulos, P. Prokopidis, V. Papavasileiou, A. Katsamanis, S. Piperidis, and V. Katsouros, "Meltemi: The first open large language model for greek," 2024. [Online]. Available: https://arxiv.org/abs/2407.20743

[19] N. Sengupta, S. K. Sahu, B. Jia, S. Katipomu, H. Li, F. Koto, W. Marshall, G. Gosal, C. Liu, Z. Chen, O. M. Afzal, S. Kamboj, O. Pandit, R. Pal, L. Pradhan, Z. M. Mujahid, M. Baali, X. Han, S. M. Bsharat, A. F. Aji, Z. Shen, Z. Liu, N. Vassilieva, J. Hestness, A. Hock, A. Feldman, J. Lee, A. Jackson, H. X. Ren, P. Nakov, T. Baldwin, and E. Xing, "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023. [Online]. Available: https://arxiv.org/abs/2308.16149

[20] D. Hernandez, T. Brown, T. Conerly, N. DasSarma, D. Drain, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume, S. Johnston, B. Mann, C. Olah, C. Olsson, D. Amodei, N. Joseph, J. Kaplan, and S. McCandlish, "Scaling laws and interpretability of learning from repeated data," 2022. [Online]. Available: https://arxiv.org/abs/2205.10487

[21] T. Nguyen, C. V. Nguyen, V. D. Lai, H. Man, N. T. Ngo, F. Dernoncourt, R. A. Rossi, and T. H. Nguyen, "Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages," 2023. [Online]. Available: https://arxiv.org/abs/2309.09400

[22] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," 2021. [Online]. Available: https://arxiv.org/abs/2010.11934

[23] M. Ali, M. Fromm, K. Thellmann, J. Ebert, A. A. Weber, R. Rutmann, C. Jain, M. Lübbering, D. Steinigen, J. Leveling, K. Klug, J. S. Buschhoff, L. Jurkschat, H. Abdelwahab, B. J. Stein, K.-H. Sylla, P. Denisov, N. Brandizzi, Q. Saleem, A. Bhowmick, L. Helmer, C. John, P. O. Suarez, M. Ostendorff, A. Jude, L. Manjunath, S. Weinbach, C. Penke, O. Filatov, S. Asaadi, F. Barth, R. Sifa, F. Küch, A. Herten, R. Jäkel, G. Rehm, S. Kesselheim, J. Köhler, and N. Flores-Herr, "Teuken-7b-base & teuken-7b-instruct: Towards european llms," 2024. [Online]. Available: https://arxiv.org/abs/2410.03730

[24] O. de Gibert, G. Nail, N. Arefyev, M. Bañón, J. van der Linde, S. Ji, J. Zaragoza-Bernabeu, M. Aulamo, G. Ramírez-Sánchez, A. Kutuzov, S. Pyysalo, S. Oepen, and J. Tiedemann, "A new

massive multilingual dataset for high-performance language technologies," 2024. [Online]. Available: https://arxiv.org/abs/2403.14009

[25] D. Roussis, L. Voukoutis, G. Paraskevopoulos, S. Sofianopoulos, P. Prokopidis, V. Papavasileiou, A. Katsamanis, S. Piperidis, and V. Katsouros, "Krikri: Advancing open large language models for greek," 2025. [Online]. Available: https://arxiv.org/abs/2505.13772

[26] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[27] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[28] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating training data makes language models better," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8424--8445. [Online]. Available: https://aclanthology.org/2022.acl-long.577/

[29] K. Tirumala, D. Simig, A. Aghajanyan, and A. S. Morcos, "D4: Improving LLM pretraining via document de-duplication and diversification," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=CG0L2PFrb1

[30] A. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, 1997, pp. 21--29.

[31] A. Z. Broder, "Identifying and filtering near-duplicate documents," in *Combinatorial Pattern Matching*, R. Giancarlo and D. Sankoff, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1--10.

[32] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, ser. STOC '98. New York, NY, USA: Association for Computing Machinery, 1998, p. 604–613. [Online]. Available: https://doi.org/10.1145/276698.276876

[33] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets, 2nd Ed*. Cambridge University Press, 2014. [Online]. Available: http://www.mmds.org/

[34] J. Leveling, L. Helmer, B. Stein, D. Wegener, Z. Sheikh, E. Fernandes, and H. Abdelwahab, *Evaluation of Document Deduplication Algorithms for Large Text Corpora*. Cham Springer, 03 2025, pp. 390--404.

[35] J. Depoix, "Youtube transcript api: A python api to retrieve the transcript/subtitles for a given youtube video." 2025. [Online]. Available: https://pypi.org/project/youtube-transcript-api/

[36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[37] D. Chuan and G. Klein, "faster-whisper: reimplementation of openai's whisper model using ctranslate2, which is a fast inference engine for transformer models." 2025. [Online]. Available: https://pypi.org/project/faster-whisper/

[38] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui, "Glam: Efficient scaling of language models with mixture-of-experts," 2022. [Online]. Available: https://arxiv.org/abs/2112.06905

[39] L. Loukas, N. Smyrnioudis, C. Dikonomaki, S. Barbakos, A. Toumazatos, J. Koutsikakis, M. Kyriakakis, M. Georgiou, S. Vassos, J. Pavlopoulos, and I. Androutsopoulos, "Gr-nlp-toolkit: An open-source nlp toolkit for modern greek," 2024. [Online]. Available: https://arxiv.org/abs/2412.08520

[40] I. Evdaimon, H. Abdine, C. Xypolopoulos, S. Outsios, M. Vazirgiannis, and G. Stamou, "Greekbart: The first pretrained greek sequence-to-sequence model," 2023. [Online]. Available: https://arxiv.org/abs/2304.00869

[41] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos, "Greek-bert: The greeks visiting sesame street," in *11th Hellenic Conference on Artificial Intelligence*, ser. SETN 2020. ACM, Sep. 2020, p. 110–117. [Online]. Available: http://dx.doi.org/10.1145/3411408.3411440

[42] J. Pavlopoulos, J. Bakagianni, K. Pouli, and M. Gavriilidou, "Open or closed llm for lesser-resourced languages? lessons from greek," 2025. [Online]. Available: https://arxiv.org/abs/2501.12826

[43] A. Vogelsang and J. Fischbach, "Using large language models for natural language processing tasks in requirements engineering: A systematic guideline," 2024. [Online]. Available: https://arxiv.org/abs/2402.13823

[44] Y. Huang, K. Tang, and M. Chen, "Leveraging large language models for enhanced nlp task performance through knowledge distillation and optimized training strategies," 2024. [Online]. Available: https://arxiv.org/abs/2402.09282

[45] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[46] A. Kargaran, A. Imani, F. Yvon, and H. Schuetze, "Glotlid: Language identification for low-resource languages," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023, p. 6155–6218. [Online]. Available: http://dx.doi.org/10.18653/v1/2023.findings-emnlp.410

[47] Artifex, "Pymupdf: a high performance python library for data extraction, analysis, conversion & manipulation of pdf (and other) documents." 2025. [Online]. Available: https://pymupdf.readthedocs.io/en/latest/

[48] C. Mou, "text-dedup: a collection of text deduplication scripts." 2024. [Online]. Available: https://github.com/ChenghaoMou/text-dedup/

[49] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[50] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, "Wildchat: 1m chatgpt interaction logs in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2405.01470

[51] N. Ding, Y. Chen, B. Xu, Y. Qin, Z. Zheng, S. Hu, Z. Liu, M. Sun, and B. Zhou, "Enhancing chat language models by scaling high-quality instructional conversations," 2023. [Online]. Available: https://arxiv.org/abs/2305.14233

[52] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vazquez, and S. Virpioja, "Democratizing neural machine translation with opus-mt," 2023. [Online]. Available: https://arxiv.org/abs/2212.01936

[53] J. Pang, F. Ye, L. Wang, D. Yu, D. F. Wong, S. Shi, and Z. Tu, "Salute the classic: Revisiting challenges of machine translation in the age of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.08350

[54] C. Lyu, Z. Du, J. Xu, Y. Duan, M. Wu, T. Lynn, A. F. Aji, D. F. Wong, S. Liu, and L. Wang, "A paradigm shift: The future of machine translation lies with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2305.01181

[55] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, and Z. Tu, "Document-level machine translation with large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16 646--16 661. [Online]. Available: https://aclanthology.org/2023.emnlp-main.1036/

[56] M. Enis and M. Hopkins, "From llm to nmt: Advancing low-resource machine translation with claude," 2024. [Online]. Available: https://arxiv.org/abs/2404.13813

[57] P.-C. Langlais, C. R. Hinostroza, M. Nee, C. Arnett, P. Chizhov, E. K. Jones, I. Girard, D. Mach, A. Stasenko, and I. P. Yamshchikov, "Common corpus: The largest collection of ethical data for llm pre-training," 2025. [Online]. Available: https://arxiv.org/abs/2506.01732

[58] L. Burchell, O. de Gibert, N. Arefyev, M. Aulamo, M. Bañón, P. Chen, M. Fedorova, L. Guillou, B. Haddow, J. Hajič, J. Helcl, E. Henriksson, M. Klimaszewski, V. Komulainen, A. Kutuzov, J. Kytöniemi, V. Laippala, P. Mæhlum, B. Malik, F. Mehryary, V. Mikhailov, N. Moghe, A. Myntti, D. O'Brien, S. Oepen, P. Pal, J. Piha, S. Pyysalo, G. Ramírez-Sánchez, D. Samuel, P. Stepachev, J. Tiedemann, D. Variš, T. Vojtěchová, and J. Zaragoza-Bernabeu, "An expanded massive multilingual dataset for high-performance language technologies (hplt)," 2025. [Online]. Available: https://arxiv.org/abs/2503.10267

[59] Z. Li, Z. Yang, and M. Wang, "Reinforcement learning with human feedback: Learning dynamic choices via pessimism," 2023. [Online]. Available: https://arxiv.org/abs/2305.18438

[60] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," 2024. [Online]. Available: https://arxiv.org/abs/2305.18290

[61] F. Bianchi, M. Suzgun, G. Attanasio, P. Röttger, D. Jurafsky, T. Hashimoto, and J. Zou, "Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions," 2024. [Online]. Available: https://arxiv.org/abs/2309.07875

[62] J. Abadji, P. O. Suarez, L. Romary, and B. Sagot, "Towards a cleaner document-oriented multilingual crawled corpus," 2022. [Online]. Available: https://arxiv.org/abs/2201.06642

[63] O. Oak, "From min-hashing to locality sensitive hashing: The complete process," *Medium*, 2024. [Online]. Available: https://medium.com/@omkarsoak/from-min-hashing-to-locality-sensitive-hashing-the-complete-process-b88b298d71a1

[64] C. Twersky, "scrapetube: Scrape youtube without the official youtube api and without selenium." 2023. [Online]. Available: https://pypi.org/project/scrapetube/

[65] "Yt-dlp: a feature-rich command-line audio/video downloader with support for thousands of sites." 2025. [Online]. Available: https://github.com/yt-dlp/yt-dlp/

[66] X.-Y. Fu, C. Chen, M. T. R. Laskar, S. B. TN, and S. Corston-Oliver, "Improving punctuation restoration for speech transcripts via external data," 2021. [Online]. Available: https://arxiv.org/abs/2110.00560

[67] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 10 2007, pp. 629 -- 633.

[68] G. Penedo, H. Kydlíček, L. B. allal, A. Lozhkov, M. Mitchell, C. Raffel, L. V. Werra, and T. Wolf, "The fineweb datasets: Decanting the web for the finest text data at scale," 2024. [Online]. Available: https://arxiv.org/abs/2406.17557

[69] Y. Tokpanov, B. Millidge, P. Glorioso, J. Pilault, A. Ibrahim, J. Whittington, and Q. Anthony, "Zyda: A 1.3t dataset for open language modeling," 2024. [Online]. Available: https://arxiv.org/abs/2406.01981

[70] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," 2022. [Online]. Available: https://arxiv.org/abs/2207.04672

[71] S. Jain, "tiktoken: a fast bpe tokeniser for use with openai's models." 2025. [Online]. Available: https://pypi.org/project/tiktoken/