



# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΜΣ «ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΥΠΗΡΕΣΙΕΣ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ  
ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ»

**Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση  
εργαλείων από  
διαλογικά συστήματα τεχνητής νοημοσύνης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Της

ΣΙΩΖΙΟΥ ΚΥΡΙΑΚΗΣ  
Α.Μ. 2117

**Επιβλέπων :** Τζήμας Ιωάννης

**Μέλη εξεταστικής επιτροπής:**

Πάτρα, Ιούνιος, 2024



## Ευχαριστίες

Ολοκληρώνοντας αυτή την εργασία, θα ήθελα να ευχαριστήσω θερμά τους καθηγητές μου κ. Τζήμα Ιωάννη και Ζέρβα Παναγιώτη για την καθοδήγηση, τις συμβουλές και την συνεργασία τους το διάστημα της εκπόνησης της εργασίας. Ήταν μια πορεία με ενδιαφέροντες σταθμούς και εμπειρίες που θα με βοηθήσει και στη συνέχεια της επαγγελματικής μου ζωής.

Επίσης, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου για τη συμπαράσταση και την υπομονή τους. Ήταν και είναι μια πηγή δύναμης.

© 2024

της

Σιώζιου Κυριακής

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Πανεπιστήμιο Πελοποννήσου



## Πίνακας περιεχομένων

<b>1</b>	<b>Εισαγωγή</b>	<b>1</b>
1.1	Ανάλυση αγοράς εργασίας και οι εξελίξεις στην επεξεργασία δεδομένων	1
1.2	Αντικείμενο διπλωματικής	3
1.3	Δομή της διπλωματικής	3
<b>2</b>	<b>Θεωρητικό πλαίσιο</b>	<b>5</b>
2.1	Συνομιλητική Τεχνητή Νοημοσύνη (Conversational AI)	5
2.2	Επεξεργασία Φυσικής Γλώσσας (NLP)	6
2.2.1	<i>Περιπτώσεις χρήσης NLP</i>	8
2.3	Αναπαράσταση κειμένου	9
2.3.1	Βασικές προσεγγίσεις διανυσματοποίησης	10
2.3.2	<i>Κατανεμημένες αναπαραστάσεις</i>	12
2.3.3	<i>Contextual Embeddings</i>	13
2.4	Κατηγοριοποίηση κειμένου (text classification)	13
2.5	Εξαγωγή Πληροφορίας (Information Extraction)	14
2.5.1	<i>Εφαρμογές Εξαγωγής Πληροφορίας</i>	14
2.5.2	<i>Keyphrase extraction (εξαγωγή φράσεων-κλειδιών)</i>	16
2.5.3	<i>Name Entity Recognition (NER)</i>	17
2.6	LLMs (Large Language Models – Μεγάλα Γλωσσικά Μοντέλα)	18
2.6.1	<i>Αρχιτεκτονική Transformer</i>	19
2.6.2	<i>Δυνατότητες των Μεγάλων Γλωσσικών Μοντέλων</i>	28
2.7	Prompt engineering	31
2.7.1	<i>Ρυθμίσεις των LLMs</i>	32
2.7.2	<i>Βασικά στοιχεία του prompt engineering</i>	32
2.7.3	<i>Γενικές οδηγίες/στρατηγικές για την δημιουργία προτροπών</i>	33
2.7.4	<i>Τεχνικές δημιουργίας προτροπών (prompts)</i>	34
2.8	Περιοχές εφαρμογής	35
2.9	Ανάλυση Αγοράς Εργασίας	37
<b>3</b>	<b>Μεθοδολογία</b>	<b>38</b>
3.1	Συλλογή Δεδομένων	39
3.2	Επιλεγμένα μοντέλα και εφαρμογές	40
3.2.1	<i>GPT-3.5-turbo</i>	40
3.2.2	<i>Llama-2-chat</i>	42

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης	
3.2.3 <i>Mistral 7B</i> .....	42
3.2.4 <i>Mini orca</i> .....	42
3.2.5 <i>GPT4All</i> .....	43
3.3 Υλοποίηση .....	43
<b>4 Αποτελέσματα και Ανάλυση</b> .....	<b>46</b>
4.1 Περιγραφή των αποτελεσμάτων .....	46
<b>5 Συζήτηση και Μελλοντική Έρευνα</b> .....	<b>55</b>
5.1 Προτάσεις για επέκταση της εργασίας.....	56
<b>Βιβλιογραφία</b> .....	<b>59</b>

## Λίστα Σχημάτων

Σχήμα 1: Πως συνδέονται η Επεξεργασία Φυσικής Γλώσσας (NLP), η Μηχανική Μάθηση (ML) και η Βαθιά Μάθηση (DL) [σχήμα από 2] .....	7
Σχήμα 2: Ετικέτες.....	14
Σχήμα 3: Άρθρο.....	16
Σχήμα 4: Βασικές συνιστώσες των LLMs [σχήμα από 6].....	18
Σχήμα 5: Αρχιτεκτονική Transformer [σχήμα από 5].....	19
Σχήμα 6: Κωδικοποιητές - Αποκωδικοποιητές Transformer [σχήμα από 4] .....	20
Σχήμα 7: Tokenization [σχήμα από 81].....	20
Σχήμα 8: Embeddings [σχήμα από 81].....	21
Σχήμα 9: Token Embeddings [σχήμα από 8] .....	22
Σχήμα 10: Self-Attention Softmax [σχήμα από 4] .....	23
Σχήμα 11: GPT-2 , BERT Αρχιτεκτονική Transformer [σχήμα από 12].....	25
Σχήμα 12: Encoder Block – Transformer [σχήμα από 4].....	26
Σχήμα 13: Decoder Block – Transformer [σχήμα από 4].....	26
Σχήμα 14: Self-Attention and Masked Self-Attention [σχήμα από 12].....	27
Σχήμα 15: LLM Transfer Learning [σχήμα από 10] .....	28
Σχήμα 16:Βασική Προτροπή.....	31
Σχήμα 17 – Παράδειγμα προτροπής με βάση την τεχνική Chain-of-Thought [σχήμα από 51] .....	35
Σχήμα 18 Ποσοστό έγκυρων εξόδων ανά μεγάλο γλωσσικό μοντέλο .....	46
Σχήμα 19 Κατανομή για Ομαδοποιημένο Job Title (Τίτλος Θέσης).....	47
Σχήμα 20 Κατανομή για Ομαδοποιημένο Industry (κλάδος βιομηχανίας) .....	47
Σχήμα 21 Κατανομή για Ομαδοποιημένο Location (Τοποθεσία) .....	48
Σχήμα 22 Σύννεφο-λέξεων για τις ήπιες δεξιότητες που επέστρεψε το μοντέλο Llama.....	48
Σχήμα 23 Σύννεφο-λέξεων για τις ήπιες δεξιότητες που επέστρεψε το μοντέλο GPT3.5 Turbo.....	49
Σχήμα 24 Σύννεφο-λέξεων για τις ήπιες δεξιότητες που επέστρεψε το μοντέλο Mistral .....	49
Σχήμα 25 Σύννεφο-λέξεων για τις ήπιες δεξιότητες που επέστρεψε το μοντέλο Orca .....	50
Σχήμα 26 Σύννεφο-λέξεων για τις σκληρές δεξιότητες που επέστρεψε το μοντέλο Llama.....	50
Σχήμα 27 Σύννεφο-λέξεων για τις σκληρές δεξιότητες που επέστρεψε το μοντέλοGPT 3.5 Turbo.....	51
Σχήμα 28 Σύννεφο-λέξεων για τις σκληρές δεξιότητες που επέστρεψε το μοντέλο Mistral .....	51
Σχήμα 29 Σύννεφο-λέξεων για τις σκληρές δεξιότητες που επέστρεψε το μοντέλο Orca .....	52
Σχήμα 30 Cosine Similarity.....	52
Σχήμα 31 Similarity Scores ανά μοντέλο .....	53



Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Σχήμα 32. Function Calling [53] .....57

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

## Περίληψη

Η πρόσφατη πρόοδος στα Μεγάλα Γλωσσικά Μοντέλα έχει ανοίξει νέες δυνατότητες για την χρήση τους σε διάφορους τομείς. Αυτή η εργασία επικεντρώνεται στη διερεύνηση των δυνατοτήτων των LLM στη δομημένη εξαγωγή πληροφοριών από αδόμητο κείμενο, ειδικά αναφορικά με τις αγγελίες θέσεων εργασίας. Συγκρίνουμε εμπορικά και ανοιχτού κώδικα Μεγάλα Γλωσσικά Μοντέλα για να δούμε πόσο καλά μπορούν να αντλήσουν βασικές πληροφορίες από ένα σύνολο αγγελιών εργασίας από τον τουριστικό τομέα της Ελλάδας. Στόχος μας είναι να κατανοήσουμε τις διαφορές απόδοσης μεταξύ αυτών των μοντέλων και να αξιολογήσουμε τη γενική τους εφαρμογή σε εργασίες εξαγωγής πληροφοριών πραγματικού κόσμου. Επιδιώκουμε να αξιολογήσουμε και να συγκρίνουμε την ικανότητα αυτών των μοντέλων στον ακριβή εντοπισμό και την εξαγωγή συγκεκριμένων σημείων δεδομένων όπως ο τίτλος εργασίας, η εταιρεία, η βιομηχανία, η τοποθεσία, οι δεξιότητες και οι δεξιότητες. Απώτερος στόχος είναι η εργασία αυτή να συμβάλλει στην κατανόησή μας για το πόσο αποτελεσματικά είναι τα LLM σε εργασίες εξαγωγής πληροφοριών στον πραγματικό κόσμο και να υπογραμμίσει τις διαφορές στην απόδοση μεταξύ διαφόρων μοντέλων τελευταίας τεχνολογίας.

**Λέξεις Κλειδιά:** Μεγάλα Γλωσσικά Μοντέλα, Επεξεργασία Φυσικής Γλώσσας, Ανάλυση αγοράς εργασίας, Εξαγωγή Πληροφορίας, Διαδικτυακά δεδομένα

## Abstract

The recent progress in Large Language Models has opened up new possibilities for their application in different domains. This work focuses on exploring the potential of LLMs in structured information extraction, specifically in the context of job postings. We compare commercial and open-source LLMs to see how well they can extract key information from job postings in Greece's tourism sector. Our goal is to understand the performance differences between these models and assess their general applicability in real-world information extraction tasks. We aim to evaluate and compare the capability of these models in accurately identifying and extracting specific data points such as Job Title, Company, Industry, Location, Soft Skills, and Hard Skills. This research contributes to our understanding of how practical LLMs are in real-world information extraction tasks and highlights the differences in performance among various state-of-the-art models.

**Keywords:** *Large Language Models, Natural Language Processing, Labor Market Analysis, Information Extraction, Online data.*

# 1

## *Εισαγωγή*

### *1.1 Ανάλυση αγοράς εργασίας και οι εξελίξεις στην επεξεργασία δεδομένων*

Στο πλαίσιο του σύγχρονου ψηφιακού τοπίου, η διαχείριση και η αξιοποίηση δεδομένων από πολλαπλές πηγές, ιδιαίτερα σε σχέση με την αγορά εργασίας, είναι μια σημαντική πρόκληση. Οι αγγελίες εργασίας, μια σημαντική και συχνά υπό-εκμεταλλευμένη πηγή, παρέχουν πολύτιμες πληροφορίες για τη δυναμική της αγοράς. Οι συμβατικές τεχνικές που χρησιμοποιούνται για την ανάλυση της αγοράς εργασίας, ως επί το πλείστον χειροκίνητες ή ημιαυτόματες, εμφανίζουν περιορισμούς όσον αφορά στην επεκτασιμότητα, στην ταχύτητα και στο βάθος της ανάλυσης. Η στροφή προς σύγχρονες μεθόδους επεξεργασίας δεδομένων, ειδικά στην οργάνωση και κατηγοριοποίηση των καταχωρίσεων θέσεων εργασίας, είναι καθοριστική για την αποκάλυψη πολύτιμων ενδείξεων, όπως οι τάσεις στις δεξιότητες και οι οικονομικές μεταβολές. Οι εξελίξεις στην ανάλυση δεδομένων σηματοδοτούν μια στροφή προς πιο προηγμένα εργαλεία, μεταβαίνοντας από τη χειροκίνητη ανάλυση βιογραφικών σε μεθόδους βασισμένες σε λογισμικό, τονίζοντας έτσι τον μετασχηματισμό στον τρόπο κατανόησης των τάσεων της αγοράς εργασίας [23].

Η μετάβαση από τις συμβατικές μεθόδους ανάλυσης της αγοράς εργασίας στην επεξεργασία φυσικής γλώσσας (NLP) αντιπροσωπεύει μια σημαντική εξέλιξη. Παρόλο που οι παραδοσιακές μέθοδοι αποτελούν την βάση, η αποτελεσματικότητά τους περιορίζεται από την επιχειρησιακή τους ικανότητα και το περιορισμένο αναλυτικό τους εύρος. Παρ' όλη τη χρησιμότητά τους, οι προσεγγίσεις αυτές, όπως η συλλογή δεδομένων μέσω στατιστικών εργαλείων και οικονομικών μετρήσεων ή η ανάλυση οικονομικών δεικτών όπως η ανεργία, αντιμετωπίζουν προκλήσεις στην

κατανόηση των συνεχώς μεταβαλλόμενων τάσεων της αγοράς εργασίας και των λεπτών αποχρώσεων αυτών. Οι μέθοδοι αυτές είναι συχνά χρονοβόρες και μπορεί να μην αναδεικνύουν τις υποκείμενες αιτίες ή τις προοπτικές που διαμορφώνουν τις μελλοντικές αγορές εργασίας.

Σε αντίθεση, η Επεξεργασία Φυσικής Γλώσσας (NLP) προσφέρει τη δυνατότητα να αναλύσει τεράστιους όγκους από κείμενα και δεδομένα γλωσσικής φύσης με ταχύτητα και ακρίβεια, ανακαλύπτοντας πρότυπα και συσχετίσεις που δεν είναι προφανή με τις παραδοσιακές μεθόδους. Αυτή η αναλυτική δύναμη επιτρέπει την ανίχνευση των εξελίξεων στις απαιτήσεις δεξιοτήτων, στις αλλαγές στις γεωγραφικές συγκεντρώσεις των θέσεων εργασίας και τις διακυμάνσεις στην προσφορά εργασίας σε σχέση με οικονομικές κρίσεις ή τεχνολογικές καινοτομίες. Η NLP μπορεί να θεωρηθεί ως ένας καταλύτης που ενισχύει και επιταχύνει τις αναλυτικές δυνατότητες στην αγορά εργασίας, παρέχοντας μια πιο σφαιρική και διεισδυτική ματιά στις τρέχουσες και μελλοντικές συνθήκες. Οι εξελίξεις στην NLP έχουν παίξει καθοριστικό ρόλο στην υπέρβαση των προαναφερθέντων περιορισμών, μετασχηματίζοντας τον τρόπο που αναλύουμε και ερμηνεύουμε τις αγορές εργασίας. Μέσω της αυτοματοποίησης της συλλογής και επεξεργασίας γλωσσικών δεδομένων, η NLP επιτρέπει την ανάλυση μεγάλων όγκων κειμένων, όπως είναι οι αγγελίες εργασίας, τα βιογραφικά σημειώματα και οι επαγγελματικές αξιολογήσεις, αναδεικνύοντας με σαφήνεια τις ανάγκες και τις τάσεις της αγοράς. Αυτή η τεχνολογία επιτρέπει επίσης την ανίχνευση και ανάλυση του συναισθηματικού τόνου των κειμένων, παρέχοντας πολύτιμες πληροφορίες αναφορικά με τη διάθεση και τις αντιλήψεις των εργαζομένων, την ποιότητα των εργασιακών περιβαλλόντων. Η αυξανόμενη ενσωμάτωση της NLP στην ανάλυση αγοράς εργασίας δεν καθιστά μόνο τη διαδικασία πιο αποδοτική, αλλά και περισσότερο προσαρμοσμένη στις δυναμικά αλλαγμένες συνθήκες της σύγχρονης οικονομικής σκηνής. Η ενσωμάτωση της μηχανικής μάθησης, ιδιαίτερα των τεχνικών βαθιάς μάθησης, έχει επιφέρει σημαντικό μετασχηματισμό στην NLP [63] βελτιώνοντας την ικανότητά της να εκτελεί πιο σύνθετες εργασίες όπως η πρόβλεψη μελλοντικών τάσεων της αγοράς και η διαχείριση μεγάλων συνόλων δεδομένων. Αυτό κατέστησε δυνατή και την άμεση μάθηση από δεδομένα [42] εξελίσσοντας περαιτέρω τις δυνατότητές της. Έτσι, η είσοδος της NLP, ιδιαίτερα με τη δημιουργία της αρχιτεκτονικής μετασχηματιστή (transformer) [5] και των Μεγάλων Γλωσσικών Μοντέλων (LLMs), έφερε αυξημένη αποτελεσματικότητα και βαθύτερη κατανόηση της αγοράς [64]. Τα LLMs είναι προηγμένα συστήματα τεχνητής νοημοσύνης που εκπαιδεύονται με μεγάλες ποσότητες δεδομένων κειμένου για να παράγουν κείμενο που μιμείται τον ανθρώπινο τρόπο γραφής, παρέχοντας συχνά εντυπωσιακές δυνατότητες σε ποικίλες γλωσσικές εργασίες. Ιδιαίτερα αυτά που είναι προσαρμοσμένα για αναλυτικές εργασίες, έχουν επιδείξει προηγμένες ικανότητες σε τομείς που αποδείχθηκαν πρόκληση για τις συμβατικές προσεγγίσεις, όπως της περίληψης, της μετάφρασης και της ερμηνείας με βάση τα συμφραζόμενα (contextual interpretation) [15, 68]. Επιπλέον, διαθέτουν την ικανότητα ανάλυσης και παραγωγής απαντήσεων που εμβαθύνουν σε σύνθετα ζητήματα, αυξάνοντας τη χρησιμότητά τους σε επαγγελματικά περιβάλλοντα και έρευνα. Η εισαγωγή LLM μοντέλων, όπως το GPT-3 σηματοδοτεί μια σημαντική πρόοδο στην χρήση της τεχνητής νοημοσύνης για την επεξεργασία γλώσσας και την ανάλυση της αγοράς εργασίας [25,43].

## **1.2 Αντικείμενο διπλωματικής**

Ο στόχος αυτής της μελέτης είναι να διερευνήσει πως η χρήση των Μεγάλων Γλωσσικών Μοντέλων επηρεάζει την εξαγωγή πληροφοριών από αγγελίες εργασίας. Μελετώντας ένα σύνολο δεδομένων από αγγελίες θέσεων εργασίας, συλλεγμένες από διάφορες διαδικτυακές πλατφόρμες, ιδιαίτερα από την τουριστική βιομηχανία της Ελλάδας, επιδιώκουμε να προσδιορίσουμε πως τα LLMs μπορούν να συμβάλλουν σημαντικά στην ανάλυση της αγοράς εργασίας. Αυτή η μέθοδος διευκολύνει μια πιο λεπτομερή και πολυδιάστατη ανάλυση της δυναμικής της αγοράς εργασίας, προσφέροντας πλεονεκτήματα έναντι των παραδοσιακών προσεγγίσεων.

Η ενσωμάτωση των μεγάλων γλωσσικών μοντέλων στην ανάλυση της αγοράς εργασίας σηματοδοτεί μια εξέλιξη με μεταμορφωτική δυναμική στη μελέτη των τάσεων της απασχόλησης. Αξιοποιώντας τις προηγμένες δυνατότητες επεξεργασίας φυσικής γλώσσας των LLMs, οι ερευνητές είναι σε θέση να αναλύουν και να ερμηνεύουν πολύπλοκα σύνολα δεδομένων από αγγελίες εργασίας με βάθος και ακρίβεια που δεν ήταν εφικτή με τις προηγούμενες μεθόδους. Αυτή η καινοτόμος προσέγγιση όχι μόνο διευρύνει το αναλυτικό πεδίο εφαρμογής, αλλά επίσης ενισχύει την προγνωστική ακρίβεια των μοντέλων τεχνητής νοημοσύνης σχετικά με τη δυναμική της αγοράς εργασίας. Τα LLMs, χάρη στην ικανότητά τους να επεξεργάζονται και να 'καταλαβαίνουν' τεράστιους όγκους φυσικής γλώσσας, αναδεικνύουν τις μεταβαλλόμενες απαιτήσεις για δεξιότητες, τις εργασιακές τάσεις και τις περιοχές με αυξημένη ζήτηση για θέσεις εργασίας.

Η τρέχουσα εργασία προσπαθεί να αναπτυχθεί σε αυτό το αναπτυσσόμενο πεδίο, διενεργώντας μια ολοκληρωμένη αξιολόγηση διαφόρων LLMs, ως προς την αποτελεσματικότητά τους στην επεξεργασία και ανάλυση των δεδομένων από αγγελίες εργασίας. Συγκρίνουμε την απόδοση τεσσάρων LLMs ως προς την εξαγωγή δομημένης πληροφορίας από ένα σύνολο από αγγελίες εργασίας. Μέσω αυτής της ανάλυσης, η μελέτη στοχεύει να δώσει μια διαφοροποιημένη κατανόηση των τάσεων της αγοράς εργασίας, διευκολύνοντας έτσι τις διαδικασίες λήψης αποφάσεων βασιζόμενες σε περισσότερες πληροφορίες. Με αυτόν τον τρόπο, συμβάλλει στη βιβλιογραφία για την ανάλυση της αγοράς εργασίας, προσφέροντας νέες γνώσεις και μεθοδολογίες που επισημαίνουν τις δυνατότητες των LLMs στην εξέλιξη των προσεγγίσεων που βασίζονται σε δεδομένα για την κατανόηση των τάσεων απασχόλησης και την πρόγνωση των οικονομικών αλλαγών.

## **1.3 Δομή της διπλωματικής**

Η δομή της εργασίας έχει ως εξής: Στο κεφάλαιο 2 παραθέτουμε το θεωρητικό πλαίσιο γύρω από το θέμα μας. Αναλυτικότερα, εξετάζουμε το τι είναι συνομιλητική τεχνητή νοημοσύνη και πως χρησιμοποιείται σε εφαρμογές που συναντάμε καθημερινά. Κατόπιν, αναλύουμε την επεξεργασία φυσικής γλώσσας, την εξέλιξή της στον χρόνο και τις περιπτώσεις χρήσης της. Ακολουθεί η παρουσίαση των επιλογών για την αναπαράσταση κειμένου, η έννοια της κατηγοριοποίησης κειμένου και αναλύεται η εξαγωγή πληροφορίας και οι εφαρμογές της. Στη συνέχεια παρουσιάζονται τα Μεγάλα Γλωσσικά Μοντέλα (LLMs), οι δυνατότητές τους, καθώς και η αρχιτεκτονική transformer που έπαιξε καθοριστικό ρόλο στην έκρηξη της εξάπλωσής τους

τα τελευταία χρόνια, όπως επίσης και οι περιοχές εφαρμογής των LLMs. Τέλος, παρουσιάζεται μια επισκόπηση του prompt engineering, δηλαδή της μελέτης του τρόπου αλληλεπίδρασης με τα LLM μέσω φυσικής γλώσσας. Στο κεφάλαιο 3, παραθέτουμε τον τρόπο σχεδίασης και υλοποίησης της εργασίας, με την παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε, των μοντέλων που επιλέχθηκαν και των ενεργειών για την εξαγωγή της ζητούμενης πληροφορίας. Τέλος, στο κεφάλαιο 4 παρουσιάζονται και αναλύονται τα αποτελέσματα της εργασίας και τα συμπεράσματα που προκύπτουν από αυτά, ενώ στο κεφάλαιο 5 προτείνονται πιθανοί τρόποι και δρόμοι για μελλοντική έρευνα στον τομέα.



# 2

## *Θεωρητικό πλαίσιο*

### *2.1 Συνομιλητική Τεχνητή Νοημοσύνη (Conversational AI)*

Η συνομιλητική Τεχνητή Νοημοσύνη είναι ένας τομέας της Τεχνητής Νοημοσύνης και ασχολείται με AI (Artificial Intelligence) πράκτορες που βασίζονται είτε στον λόγο (speech-based), είτε στο κείμενο (text-based) και έχουν την ικανότητα να προσομοιώνουν και να αυτοματοποιούν την ανθρώπινη συνομιλία και τις λεκτικές αλληλεπιδράσεις.

Οι πράκτορες συνομιλητικής AI, όπως τα chatbots, έχουν αναπτυχθεί σημαντικά το τελευταίο έτος, ειδικά λόγω δύο κύριων εξελίξεων. Πρώτον, η πρόοδος στην επίτευξη υψηλότερης υπολογιστικής ισχύος οδήγησε σε αυξημένο ερευνητικό ενδιαφέρον, επιτρέποντας την αξιοποίηση του τεράστιου όγκου δεδομένων που είναι διαθέσιμα στο διαδίκτυο και την εφαρμογή μεθόδων απαραίτητων για την ανάπτυξη μοντέλων τεχνητής νοημοσύνης, όπως η Μηχανική Μάθηση (Machine Learning) και η Βαθιά Μάθηση (Deep Learning). Δεύτερον, αποτελούν ιδανικές εφαρμογές για χρήση σε ποικίλους τομείς, όπως η υγειονομική περίθαλψη, το ηλεκτρονικό εμπόριο, η εξυπηρέτηση πελατών, ο τουρισμός και η εκπαίδευση, οι οποίοι εξαρτώνται σημαντικά από τις συνομιλίες σε φυσική γλώσσα για τις καθημερινές τους λειτουργίες. Αυτή η απότομη αύξηση της ζήτησης έχει αντιμετωπιστεί με εξίσου εντυπωσιακό ρυθμό έρευνας και ανάπτυξης, με καινοτομίες που προτείνονται συνεχώς. [1]

Η συνομιλητική Τεχνητή Νοημοσύνη επιτρέπει στους υπολογιστές να κατανοούν, να επεξεργάζονται και να ανταποκρίνονται σε συνομιλίες κειμένου ή φωνής. Επιτρέπει τη φυσική και διαδραστική επικοινωνία μεταξύ ανθρώπων και μηχανών, χρησιμοποιώντας γραπτή ή προφορική ανθρώπινη γλώσσα. Μέσω μηχανικής μάθησης, κατανόησης γλώσσας και διαχείρισης γνώσης, τα συστήματα τεχνητής νοημοσύνης συνομιλίας μπορούν να εκπαιδευτούν ώστε να

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

κατανοούν ερωτήματα, να απαντούν με ακρίβεια, ακόμη και να προβλέψουν και να προτείνουν πληροφορίες με βάση το ιστορικό συνομιλιών.

Προερχόμενη από την ιδέα της δημιουργίας ευφών συνομιλητών ικανών για ανθρώπινη αλληλεπίδραση, η συνομιλητική Τεχνητή Νοημοσύνη έχει επαναπροσδιορίσει την ψηφιακή επικοινωνία.

Μερικά παραδείγματα εφαρμογών τεχνολογίας τεχνητής νοημοσύνης συνομιλίας:

**Πράκτορες Generative AI (Παραγωγικής Τεχνητής Νοημοσύνης)** : αυτοί οι εικονικοί πράκτορες χρησιμοποιούν Generative AI για να τροφοδοτούν τις συνομιλίες κειμένου ή φωνής. Στην κατηγορία αυτή ανήκει το ChatGPT.

**Chatbots:** χρησιμοποιούνται συχνά σε εφαρμογές εξυπηρέτησης πελατών για να απαντούν σε ερωτήσεις και να παρέχουν υποστήριξη.

**Εικονικοί βοηθοί:** συχνά ενεργοποιούνται με φωνή και μπορούν να χρησιμοποιηθούν σε κινητές συσκευές και έξυπνα ηχεία

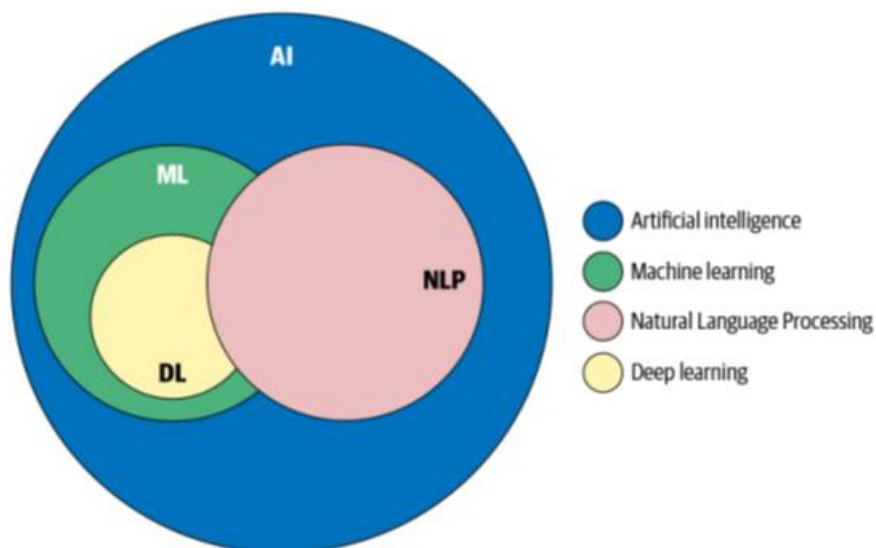
**Λογισμικό μετατροπής κειμένου σε ομιλία:** χρησιμοποιείται για τη δημιουργία ηχητικών βιβλίων ή προφορικών οδηγιών

**Λογισμικό αναγνώρισης ομιλίας:** χρησιμοποιείται για τη μεταγραφή διαλέξεων, τηλεφωνικών κλήσεων, λεζάντες και άλλα.

## 2.2 Επεξεργασία Φυσικής Γλώσσας (NLP)

Για να είναι εφικτή η δημιουργία των συστημάτων συνομιλητικής νοημοσύνης που αναφέρθηκαν παραπάνω, απαιτείται η κατανόηση, ανάλυση και επεξεργασία από το υπολογιστικό σύστημα της φυσικής γλώσσας, της γλώσσας δηλαδή που χρησιμοποιούν οι άνθρωποι. Εδώ υπεισέρχεται ο τομέας της Επεξεργασίας Φυσικής Γλώσσας (NLP), ο οποίος αποτελεί κομμάτι της Τεχνητής Νοημοσύνης.

Η σχέση μεταξύ των διαφόρων τομέων της AI φαίνεται στο παρακάτω σχήμα:



## Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Σχήμα 1: Πως συνδέονται η Επεξεργασία Φυσικής Γλώσσας (NLP), η Μηχανική Μάθηση (ML) και η Βαθιά Μάθηση (DL) [σχήμα από 2]

Η δυσκολία του εγχειρήματος για την κατανόηση και επεξεργασία της Φυσικής Γλώσσας από υπολογιστικό σύστημα είναι μεγάλη, λόγω της πολυπλοκότητας και της ασάφειάς της.

Ασάφεια σημαίνει αβεβαιότητα νοήματος. Οι περισσότερες ανθρώπινες γλώσσες είναι εγγενώς ασαφείς. Για παράδειγμα, η φράση «Τον χτύπησε κατακέφαλα» μπορεί να έχει πολλαπλές έννοιες. Π.χ. κυριολεκτικά μιλώντας, ένα αντικείμενο να τον χτύπησε ή μεταφορικά, να συνειδητοποίησε την αλήθεια. Ποια από τις δύο έννοιες ισχύει εξαρτάται από το πλαίσιο που εμφανίζεται η πρόταση. Το σημαντικό αυτό πρόβλημα ήταν ιδιαίτερα εμφανές στα πρώτα στάδια των εφαρμογών αυτόματης μετάφρασης. Επιπλέον, σημαντικό ζήτημα για το επίπεδο δυσκολίας του εγχειρήματος είναι και ο μεγάλος αριθμός φυσικών γλωσσών, με διαφορές σημαντικές σε συντακτικό, λεξιλόγιο, ιδιωματισμούς κ.λπ., καθώς και φυσικά η δημιουργικότητα της ανθρώπινης φύσης η οποία και δεν περιορίζεται από προκαθορισμένους κανόνες, όπως συμβαίνει για παράδειγμα στις γλώσσες προγραμματισμού.

Η επεξεργασία φυσικής γλώσσας (NLP) αναφέρεται στον κλάδο της τεχνητής νοημοσύνης που ενδιαφέρεται να δώσει στους υπολογιστές την ικανότητα να κατανοούν το κείμενο και τις προφορικές λέξεις με τρόπο παρόμοιο με αυτόν των ανθρώπων.

Η NLP συνδυάζει την υπολογιστική γλωσσολογία (computational linguistics—rule-based modeling of human language), την μοντελοποίηση δηλαδή της ανθρώπινης γλώσσας βάσει κανόνων, με μοντέλα στατιστικής, μηχανικής μάθησης και βαθιάς μάθησης. Αυτός ο συνδυασμός τεχνολογιών επιτρέπει στους υπολογιστές να επεξεργάζονται την ανθρώπινη γλώσσα με τη μορφή κειμένου ή φωνητικών δεδομένων και να «κατανοούν» το πλήρες νόημά της, με την πρόθεση και το συναίσθημα του ομιλητή ή του συγγραφέα.

Η επεξεργασία φυσικής γλώσσας βρίσκεται πίσω από τα προγράμματα υπολογιστών που μεταφράζουν κείμενο από μια γλώσσα σε άλλη, ανταποκρίνονται σε προφορικές εντολές και συνοφίζουν μεγάλους όγκους κειμένου γρήγορα. Δημοφιλείς εφαρμογές NLP είναι συστήματα πλοήγησης GPS που λειτουργούν με φωνή, ψηφιακοί βοηθοί, λογισμικό υπαγόρευσης ομιλίας σε κείμενο, chatbot εξυπηρέτησης πελατών κ.α.. Ωστόσο, η NLP διαδραματίζει επίσης αυξανόμενο ρόλο στις επιχειρηματικές λύσεις που βοηθούν στον εξορθολογισμό των επιχειρηματικών λειτουργιών, στην αύξηση της παραγωγικότητας των εργαζομένων και στην απλοποίηση κρίσιμων επιχειρηματικών διαδικασιών.

Η ανθρώπινη γλώσσα είναι γεμάτη με ασάφειες που καθιστούν απίστευτα δύσκολη τη σύνταξη λογισμικού που προσδιορίζει με ακρίβεια την επιδιωκόμενη έννοια του κειμένου ή των φωνητικών δεδομένων. Ομώνυμα, ομόφωνα, σαρκασμός, ιδιωματισμοί, μεταφορές, εξαιρέσεις γραμματικής και χρήσης, παραλλαγές στη δομή των προτάσεων - αυτές είναι μερικές μόνο από τις ιδιότητες της ανθρώπινης γλώσσας που χρειάζονται χρόνια για να μάθουν οι άνθρωποι και τις οποίες πρέπει να διδαχθούν οι εφαρμογές που βασίζονται σε φυσική γλώσσα για να αναγνωρίζουν και κατανοήσουν με ακρίβεια το περιεχόμενο που του δίνεται, ώστε να είναι χρήσιμες.

Αρκετές εργασίες NLP (NLP tasks) αναλύουν τα ανθρώπινα δεδομένα κειμένου και φωνής με τρόπους που βοηθούν τον υπολογιστή να κατανοήσει τι προσλαμβάνει. Ορισμένες από αυτές τις εργασίες περιλαμβάνουν τα ακόλουθα:

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

**Η αναγνώριση ομιλίας (Speech recognition)**, που ονομάζεται επίσης **ομιλία σε κείμενο**, είναι το έργο της αξιόπιστης μετατροπής των φωνητικών δεδομένων σε δεδομένα κειμένου. Η αναγνώριση ομιλίας απαιτείται για κάθε εφαρμογή που ακολουθεί φωνητικές εντολές ή απαντά σε προφορικές ερωτήσεις. Αυτό που κάνει την αναγνώριση ομιλίας μια ιδιαίτερη πρόκληση είναι ο τρόπος που μιλάνε οι άνθρωποι—γρήγορα, μπερδεμένες λέξεις μαζί, με διαφορετική έμφαση και τονισμό, με διαφορετικές προφορές και συχνά χρησιμοποιώντας λανθασμένη γραμματική.

**Επισήμανση των μερών του λόγου (Part of speech tagging)**, που ονομάζεται επίσης γραμματική σήμανση, είναι η διαδικασία προσδιορισμού του μέρους της ομιλίας μιας συγκεκριμένης λέξης ή κομματιού κειμένου, με βάση τη χρήση και το περιεχόμενό της. Π.χ. η λέξη «ηχώ» μπορεί να είναι ρήμα (ηχώ τις σάλπιγγες) ή ουσιαστικό (Ακούω την ηχώ μου).

**Η αποσαφήνιση της έννοιας της λέξης (Word sense disambiguation)** είναι η επιλογή της σημασίας μιας λέξης με πολλαπλές σημασίες, μέσω μιας διαδικασίας σημασιολογικής ανάλυσης που καθορίζει τη λέξη που έχει περισσότερο νόημα στο δεδομένο πλαίσιο. Για παράδειγμα, η αποσαφήνιση της έννοιας της λέξης βοηθά στη διάκριση της σημασίας του ρήματος «κάνω» σε «κάνω στην άκρη» (παραμερίζω) έναντι «κάνω δουλειά» (εργάζομαι).

**Η αναγνώριση ονοματισμένων οντοτήτων (Named entity recognition)**, ή NER, προσδιορίζει λέξεις ή φράσεις ως χρήσιμες οντότητες. Το NER προσδιορίζει το "Αθήνα" ως τοποθεσία ή το "Αθηνά" ως όνομα.

**Η ανάλυση συν-αναφοράς (Co-reference resolution)** είναι εργασία του προσδιορισμού για το εάν και πότε δύο λέξεις αναφέρονται στην ίδια οντότητα. Το πιο συνηθισμένο παράδειγμα είναι ο προσδιορισμός του προσώπου ή του αντικειμένου στο οποίο αναφέρεται μια συγκεκριμένη αντωνυμία (π.χ. "αυτή" = "Μαρία"), αλλά μπορεί επίσης να περιλαμβάνει τον προσδιορισμό μιας μεταφοράς ή ενός ιδιώματος στο κείμενο (π.χ. η αρκούδα δεν είναι ζώο, αλλά ένα μεγαλόσωμο τριχωτό άτομο).

**Η ανάλυση συναισθημάτων (Sentiment analysis)** επιχειρεί να εξάγει υποκειμενικές ιδιότητες όπως διάθεση, συναισθήματα, σαρκασμός, σύγχυση, καχυποψία από το κείμενο.

**Η παραγωγή φυσικής γλώσσας (Natural language generation)** που μερικές φορές περιγράφεται ως το αντίθετο της αναγνώρισης ομιλίας ή της ομιλίας σε κείμενο, είναι το καθήκον της εισαγωγής δομημένων πληροφοριών στην ανθρώπινη γλώσσα. [3]

### 2.2.1 Περιπτώσεις χρήσης NLP

Η επεξεργασία φυσικής γλώσσας είναι η κινητήρια δύναμη πίσω από τη νοημοσύνη των μηχανών σε πολλές σύγχρονες εφαρμογές του πραγματικού κόσμου. Ακολουθούν μερικά παραδείγματα:

**Ανίχνευση ανεπιθύμητων μηνυμάτων (Spam detection):** Οι τεχνολογίες ανίχνευσης ανεπιθύμητων μηνυμάτων χρησιμοποιούν τις δυνατότητες ταξινόμησης κειμένου του NLP για τη σάρωση μηνυμάτων ηλεκτρονικού ταχυδρομείου για γλώσσα που συχνά υποδηλώνει ανεπιθύμητη αλληλογραφία ή ηλεκτρονικό ψάρεμα (phishing). Αυτοί οι δείκτες μπορεί να περιλαμβάνουν υπερβολική χρήση οικονομικών όρων, χαρακτηριστικά κακή γραμματική, απειλητική γλώσσα, ανορθόγραφα ονόματα εταιρειών και άλλα. Ο εντοπισμός ανεπιθύμητων μηνυμάτων είναι ένα από τα ελάχιστα προβλήματα NLP που οι ειδικοί θεωρούν *κατά κύριο λόγο λυμένα*.

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

**Μηχανική μετάφραση (Machine translation):** Το Google Translate είναι ένα παράδειγμα εφαρμοσμένης και ευρέως διαθέσιμης τεχνολογίας NLP. Η πραγματικά χρήσιμη αυτόματη μετάφραση περιλαμβάνει πολλά περισσότερα από την αντικατάσταση λέξεων σε μια γλώσσα με λέξεις μιας άλλης. Η αποτελεσματική μετάφραση πρέπει να αποτυπώνει με ακρίβεια το νόημα και τον τόνο της γλώσσας εισόδου και να το μεταφράζει σε κείμενο με το ίδιο νόημα και επιθυμητό αντίκτυπο στη γλώσσα εξόδου. Τα εργαλεία μηχανικής μετάφρασης έχουν σημειώσει καλή πρόοδο, όσον αφορά στην ακρίβεια.

**Εικονικοί πράκτορες και chatbots (Virtual agents and chatbots):** Εικονικοί πράκτορες όπως το Siri της Apple και το Alexa της Amazon χρησιμοποιούν την αναγνώριση ομιλίας για να αναγνωρίζουν μοτίβα στις φωνητικές εντολές και τη δημιουργία φυσικής γλώσσας ώστε να ανταποκρίνονται με την κατάλληλη ενέργεια ή χρήσιμα σχόλια. Τα chatbots εκτελούν τις παραπάνω λειτουργίες ως απόκριση σε πληκτρολογημένες καταχωρήσεις κειμένου. Κάποια από αυτά έχουν την ικανότητα να αναγνωρίζουν από τα συμφραζόμενα ενδείξεις, σχετικά με τα ανθρώπινα αιτήματα και να τα χρησιμοποιούν για να παρέχουν ακόμα καλύτερες απαντήσεις ή επιλογές με την πάροδο του χρόνου. Το επόμενο στάδιο εξέλιξης για τις εφαρμογές αυτές είναι η δυνατότητα απάντησης σε ερωτήσεις, αναμενόμενες ή μη, με παραγωγή λόγου, δηλαδή η σχετική και χρήσιμη απάντηση με δικά τους λόγια.

**Ανάλυση συναισθήματος μέσω κοινωνικής δικτύωσης (Social media sentiment analysis):** Η NLP έχει γίνει ένα απαραίτητο επιχειρηματικό εργαλείο για την αποκάλυψη κρυμμένων πληροφοριών δεδομένων από κοινωνικά μέσα δικτύωσης. Η ανάλυση συναισθήματος μπορεί να εξετάσει τη γλώσσα που χρησιμοποιείται σε αναρτήσεις, απαντήσεις, κριτικές σε μέσα κοινωνικής δικτύωσης για την εξαγωγή θέσεων και συναισθημάτων ως απάντηση σε προϊόντα, προωθήσεις και εκδηλώσεις. Τις πληροφορίες αυτές χρησιμοποιούν οι εταιρείες για τον σχεδιασμό προϊόντων, διαφημιστικών καμπανιών και αλλού.

**Σύνοψη κειμένου (Text summarization):** Η σύνοψη κειμένου χρησιμοποιεί τεχνικές NLP για να αφομοιώσει τεράστιους όγκους ψηφιακού κειμένου και να δημιουργήσει περιλήψεις και συνοψίσεις για ευρετήρια, βάσεις δεδομένων έρευνας ή απλά για αναγνώστες που δεν έχουν χρόνο ή δεν επιθυμούν να διαβάσουν το πλήρες κείμενο. Οι καλύτερες εφαρμογές περίληψης κειμένου χρησιμοποιούν σημασιολογικό συλλογισμό και παραγωγή φυσικής γλώσσας (natural language generation - NLG) για να προσθέσουν χρήσιμο πλαίσιο και συμπεράσματα στις περιλήψεις.

**Η εξαγωγή πληροφορίας (Information extraction).** Όπως είναι ενδεικτικό από το όνομα, η εργασία αυτή είναι η εξαγωγή σχετικής πληροφορίας από κείμενα, όπως παραδείγματος χάριν εξαγωγή συμβάντων ημερολογίων από μηνύματα ηλεκτρονικής αλληλογραφίας ή η εξαγωγή ονομάτων ατόμων που αναφέρονται σε μια δημοσίευση σε μέσο κοινωνικής δικτύωσης.

## 2.3 Αναπαράσταση κειμένου

Για να είναι δυνατή η χρήση των τεχνικών NLP στις εφαρμογές που αναφέρθηκαν νωρίτερα, είναι απαραίτητη η αναπαράσταση των δεδομένων κειμένου που διαθέτουμε σε μια μορφή κατανοητή από τα μοντέλα τεχνητής νοημοσύνης, μορφή η οποία θα πρέπει να είναι δομημένη, να επιτρέπει τον εντοπισμό προτύπων, την εξαγωγή πληροφοριών και την απόκτηση γνώσεων. Για τον λόγο

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

αυτό τα δεδομένα κειμένου μετατρέπονται σε διανύσματα (vectors), με τέτοιο τρόπο ώστε να αντανακλούν γλωσσικές ιδιότητες του κειμένου.

Τα χαρακτηριστικά (features) του κειμένου είναι τα στοιχεία που χρησιμοποιούν οι αλγόριθμοι για την κατανόηση και την ερμηνεία του. Μπορούν να κυμαίνονται από απλούς δείκτες, όπως το μήκος μιας πρότασης ή την παρουσία συγκεκριμένων λέξεων, έως πιο σύνθετα μοτίβα, όπως το συναίσθημα που εκφράζεται σε μια παράγραφο ή η συντακτική δομή μιας πρότασης. Η επιλογή των χαρακτηριστικών παίζει σημαντικό ρόλο στην απόδοση των μοντέλων ανάλυσης κειμένου. Τα χαρακτηριστικά μπορούν να κατηγοριοποιηθούν ευρέως σε δύο τύπους: λεξιλογικά χαρακτηριστικά, όπως πλήθος λέξεων, ετικέτες μέρους του λόγου (part-of-speech tags) και σημασιολογικά χαρακτηριστικά, όπως σημασίες λέξεων βάσει των συμφραζόμενων).

Υπάρχουν διάφορες προσεγγίσεις στην αναπαράσταση των δεδομένων, την εξαγωγή δηλαδή, ουσιαστικά των χαρακτηριστικών τους (feature extraction) με τρεις βασικές κατηγορίες [2] :

*Βασικές προσεγγίσεις διανυσματοποίησης (Basic vectorization approaches)*

*Κατανεμημένες αναπαραστάσεις (Distributed representations)*

*Contextual Embeddings (ενσωματώσεις βάσει των συμφραζόμενων)*

### **2.3.1 Βασικές προσεγγίσεις διανυσματοποίησης**

Στην προσέγγιση αυτή χρησιμοποιείται το μαθηματικό μοντέλο της απεικόνισης κειμένου με διανύσματα, γνωστό ως vector model ή vector space model. Τα διανύσματα αποτελούνται από αναγνωριστικά (identifiers), όπως, για παράδειγμα, αριθμός ευρετηρίου (index number) σε ένα λεξικό από όρους. Η χρήση των διανυσμάτων βοηθά στο να υπολογιστεί η ομοιότητα δύο κομματιών κειμένου με τον υπολογισμό της ομοιότητας συνημίτονου (cosine similarity) μεταξύ των διανυσμάτων. Το συνημίτονο των  $0^\circ$  (όμοια δηλαδή διανύσματα) είναι 1 και το συνημίτονο των  $180^\circ$  είναι  $-1$ , με το συνημίτονο να μειώνεται από  $0^\circ$  σε  $180^\circ$ .

#### **2.3.1.1 Bag of words (“σακούλα λέξεων”)**

Το μοντέλο Bag-of-words προτείνει μια προσέγγιση για την εξαγωγή των χαρακτηριστικών του κειμένου αρκετά απλή και ευέλικτη. Είναι μια αναπαράσταση κειμένου που περιγράφει την εμφάνιση των λέξεων στο έγγραφο. Ονομάζεται bag (σακούλα), καθώς δεν περιέχει καμία πληροφορία για την θέση της λέξης στο κείμενο, αλλά μόνο για την παρουσία ή μη αυτής. Για την υλοποίηση του μοντέλου χρησιμοποιείται ένα λεξικό από γνωστές λέξεις και η μέτρηση της παρουσίας αυτών στο κείμενο. Η γραμματική και η σειρά των λέξεων δεν λαμβάνεται υπόψη, μόνο ο αριθμός της εμφάνισής τους στο κείμενο [77]. Ένα απλό παράδειγμα είναι το παρακάτω:

«Το βιβλίο είναι πάνω στο τραπέζι και το μολύβι είναι κάτω από το τραπέζι.» Για την πρόταση αυτή εφαρμόζοντας το μοντέλο bag-of-words θα είχαμε:

το: 3

βιβλίο: 1

είναι: 2

πάνω: 1

στο : 1

τραπέζι: 2

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

και: 1

μολύβι: 1

κάτω : 1

από : 1

Με αλφαβητική ταξινόμηση στο λεξικό που προκύπτει (από, βιβλίο, είναι, και, κάτω, μολύβι, πάνω, στο, το, τραπέζι) η αναπαράσταση της πρότασης σε διάνυσμα θα ήταν : [1,1,2,1,1,1,1,1,3,2].

Η τεχνική έχει προφανείς περιορισμούς, όπως για παράδειγμα το μέγεθος του λεξικού. Με την αύξηση του μεγέθους των κειμένων, αυξάνεται το μέγεθος του λεξικού άρα και το μέγεθος του διανύσματος [78]. Επίσης, καθώς δεν προσφέρει καμία πληροφορία σχετικά με την σημασία των λέξεων, μπορεί να θεωρηθούν όμοιες ή κοντινές προτάσεις που νοηματικά απέχουν σημαντικά.

Υπάρχουν διάφορες παραλλαγές στο βασικό μοντέλο για να αντιμετωπιστούν προβλήματα, όπως η μη χρήση των πιο συνηθισμένων λέξεων (άρθρων -το, ο, η- , συνδέσμων – και, άρα-προθέσεων -από, με- κ.λπ.). Παρ' όλους τους περιορισμούς της όμως, παραμένει μια χρήσιμη και απλή μέθοδος και για αυτό χρησιμοποιείται ευρέως, ειδικά σε θέματα κατηγοριοποίησης κειμένου (text classification) στην NLP, καθορίζοντας π.χ. το θέμα ενός εγγράφου.

### 2.3.1.2 TF – IDF

Ένα χαρακτηριστικό της μεθόδου bag-of-words που αναφέρθηκε νωρίτερα είναι ότι δεν δίνει βάρος σε κάποια λέξη, αλλά αντίθετα τις αντιμετωπίζει ισότιμα. Στην πράξη όμως, κάποιες λέξεις έχουν ειδικό βάρος αναφορικά με την σημασία τους στο κείμενο αν και μπορεί να μην απαντώνται συχνά στο έγγραφο. Για να βελτιωθεί αυτό το ζήτημα προτάθηκε [78] η χρήση της IDF (Inverse Document Frequency – αντίστροφη συχνότητα εγγράφου) σε συνδυασμό με την TF (Term Frequency – συχνότητα όρου). Η ιδέα πίσω από αυτή την μέθοδο είναι η μείωση της σημασίας κοινών λέξεων που εμφανίζονται συχνά στο κείμενο, αλλά δεν προσθέτουν βάρος στο θέμα που διαπραγματεύεται, σε συνδυασμό με την ποσοτικοποίηση της σημασίας άλλων λέξεων.

Συγκεκριμένα, η TF μετράει πόσο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο και υπολογίζεται με τον εξής τύπο, για έναν όρο  $t$  σε ένα έγγραφο  $d$ :

$$TF(t, d) = \frac{\text{(αριθμός εμφανίσεων όρου } t \text{ στο έγγραφο } d)}{\text{(συνολικός αριθμός όρων στο έγγραφο } d)}$$

Ενώ, η IDF μετράει τη σημασία ενός όρου σε μια συλλογή εγγράφων και υπολογίζεται με τον εξής τύπο:

$$IDF(t) = \log_e \frac{\text{(συνολικός αριθμός εγγράφων στη συλλογή)}}{\text{(αριθμός εγγράφων στα οποία εμφανίζεται ο όρος } t)}$$

Το τελικό σκορ TF-IDF είναι το γινόμενο των επιμέρους σκορ, δηλαδή :  $TF-IDF \text{ score} = TF * IDF$ . Ο υπολογισμός αυτός βαθμολογίας επισημαίνει λέξεις που είναι ξεχωριστές και έχουν ιδιαίτερη σημασία σε κάποιο έγγραφο. Κατόπιν, δημιουργείται πίνακας με γραμμές τους όρους και στήλες τα έγγραφα της συλλογής. Στη βιβλιογραφία υπάρχουν διαφορετικοί τρόποι υπολογισμού της TF-IDF βαθμολογίας με μικρές διαφορές. Χρησιμοποιείται σε εφαρμογές

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

αναζήτησης και ανάκτησης πληροφορίας και σε εξαγωγές λέξεων «κλειδιά» (keyword) από κείμενα.

Η συγκεκριμένη μέθοδος, αν και βελτιώνει την πιο απλή λογική του bag-of-words, έχει αντίστοιχους περιορισμούς. Με την αύξηση του μεγέθους της συλλογής γιγαντώνεται και το πρόβλημα των διαστάσεων στα διανύσματα. Επιπλέον, ενώ υπολογίζει βαρύτητα στους όρους, δεν λαμβάνει υπόψη καθόλου το πλαίσιο στο οποίο χρησιμοποιούνται ή την σημασία τους.

### 2.3.2 Κατανεμημένες αναπαραστάσεις

Η προσέγγιση αυτή στις αναπαραστάσεις κειμένου προτάθηκε για να βελτιώσει τα προβλήματα που υπήρχαν στις προηγούμενες βασικές προσεγγίσεις διανυσματοποίησης. Πιο αναλυτικά, έχουμε μια μετάβαση από απλές και συχνά αραιές (sparse) αναπαραστάσεις (δηλαδή τα περισσότερα στοιχεία στις διανυσματικές αναπαραστάσεις είναι μηδενικά), χωρίς να έχουν ληφθεί υπόψη οι σημασιολογικές σχέσεις των λέξεων, σε αναπαραστάσεις με πυκνά (dense) διανύσματα που κωδικοποιούν και σημασιολογικές ομοιότητες και το πλαίσιο της συζήτησης. Αυτά είναι τα word embeddings.

#### 2.3.2.1 Word2Vec

Το Word2Vec είναι μια τεχνική στην οποία η αναπαράσταση των λέξεων ως διανύσματα λαμβάνει υπόψη και τη σημασία των λέξεων, όχι μόνο τη συχνότητα εμφάνισής τους στο έγγραφο. Το [79] έδειξε ότι το μοντέλο Word2Vec, χρησιμοποιώντας νευρωνικά δίκτυα, μπορεί να συλλάβει την αναλογία στην αναπαράσταση των λέξεων, όπως στο παράδειγμα:  $King - Man + Woman \approx Queen$ . Στο παράδειγμα αυτό, οι πράξεις μεταξύ των διανυσματικών αναπαραστάσεων των King, Man και Woman καταλήγουν σε ένα διάνυσμα κοντά στη διανυσματική αναπαράσταση της λέξης Queen, όπως θα περιμέναμε από τη σημασιολογική εξέταση των σχέσεων μεταξύ των λέξεων.

Το μοντέλο αντλεί τη σημασία μιας λέξης από τα συμφραζόμενά της, από τις λέξεις που εμφανίζονται κοντά της στο κείμενο. Έτσι, αν δύο διαφορετικές λέξεις εμφανίζονται συχνά, σε παρόμοιο πλαίσιο, τότε είναι πολύ πιθανό οι έννοιές τους να είναι επίσης παρόμοιες. Το Word2Vec καταλήγει να δημιουργεί έναν διανυσματικό χώρο, όπου λέξεις με παρόμοια σημασία τείνουν να συγκεντρώνονται μαζί, ενώ λέξεις με πολύ διαφορετικές έννοιες απέχουν πολύ η μία από την άλλη. Επιπλέον, τα διανύσματα που προκύπτουν από το μοντέλο Word2Vec είναι πιο πυκνά και με μικρότερες διαστάσεις, βελτιώνοντας τα αποτελέσματα προηγούμενων προσεγγίσεων.

#### 2.3.2.2 GloVe

Το GloVe είναι επίσης μοντέλο για δημιουργία word embeddings [80]. Έχει μια διαφορετική προσέγγιση από το Word2Vec. Ενώ το Word2Vec εστιάζει στις προτάσεις και τις γειτονικές λέξεις για αναγνώριση σχέσεων στο τοπικό πλαίσιο, το GloVe αναζητά πόσο συχνά εμφανίζονται ζευγάρια λέξεων σε όλη τη συλλογή εγγράφων, θεωρώντας ότι η αναλογία των πιθανοτήτων συν-εμφάνισης μεταξύ των λέξεων περιέχει σημαντικές κωδικοποιημένες πληροφορίες για τη σχέση τους. Έτσι, κατασκευάζει έναν μεγάλο πίνακα συν-εμφάνισης λέξεων από ολόκληρο τη συλλογή κειμένων, όπου κάθε στοιχείο του πίνακα αντιπροσωπεύει πόσο συχνά συνυπάρχουν λέξεις μαζί



Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

σε ένα συγκεκριμένο παράθυρο περιβάλλοντος. Οι διανυσματικές αναπαραστάσεις που προκύπτουν από το GloVe είναι πυκνές, όπως και στο Word2Vec. Η επιλογή για το ποιο από τα δύο θα χρησιμοποιηθεί εξαρτάται από τις απαιτήσεις της εργασίας και τα χαρακτηριστικά των δεδομένων.

### 2.3.3 *Contextual Embeddings*

Η επόμενη εξέλιξη στην εξαγωγή χαρακτηριστικών και την διανυσματική αναπαράσταση των λέξεων εισήχθη με την εισαγωγή της αρχιτεκτονικής transformer [5] που αναλύεται παρακάτω. Με την αρχιτεκτονική αυτή κατέστη δυνατή η αναπαράσταση των λέξεων με βάση το πλαίσιο στο οποίο βρίσκονται και όχι μόνο ως αυτόνομες λέξεις και αποτελεί ένα τεράστιο βήμα βελτίωσης. Η αριθμητική αναπαράσταση των λέξεων ως διανύσματα περιέχει, πλέον, γλωσσικές και συντακτικές πληροφορίες και νόημα και υπολογίζεται λαμβάνοντας υπόψη το ευρύτερο πλαίσιο στο οποίο χρησιμοποιούνται οι λέξεις, καταλήγοντας σε πιο ακριβή κατανόηση της γλώσσας. Για τον λόγο αυτό, τα LLMs (Large Language Models) που περιγράφονται στη συνέχεια, χρησιμοποιούνται σε πλήθος γλωσσικών εργασιών.

## 2.4 *Κατηγοριοποίηση κειμένου (text classification)*

Η ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικής αλληλογραφίας (spam) αποτελεί μία από τις πιο διαδεδομένες εφαρμογές της επεξεργασίας φυσικής γλώσσας (NLP). Αυτή η διαδικασία της ανίχνευσης spam επιτυγχάνεται μέσω της ταξινόμησης κειμένου (text classification), κατηγοριοποιώντας τα μηνύματα ως επιθυμητά ή ανεπιθύμητα. Η κατηγοριοποίηση κειμένου αφορά στην ταξινόμηση κειμένου σε μία ή περισσότερες προκαθορισμένες κατηγορίες και έχει εφαρμογή σε πολλές περιοχές όπως τα κοινωνικά δίκτυα, η υγεία και η διαφήμιση. Για παράδειγμα, στις κριτικές των καταναλωτών το text classification χρησιμοποιείται για την ανάλυση των εντυπώσεων τους από τα προϊόντα και την ανίχνευση των αναγκών τους, επιτρέποντας την βελτίωση των προϊόντων. Στην υγεία εφαρμόζεται για την ταξινόμηση των ιατρικών αρχείων σε κατηγορίες όπως διαγνώσεις, θεραπείες και ιστορικό ασθενών, με στόχο τη βελτίωση των υπηρεσιών υγείας.

Για την υλοποίηση της εργασίας υπάρχουν πολλές επιλογές που ακολουθούν την ίδια δομή. Καταρχάς, επιλέγεται ο τρόπος εξαγωγής των χαρακτηριστικών από τα δεδομένα (feature extraction) είτε π.χ. με χρήση του TF-IDF, είτε με GloVe ή Word2Vec. Στην περίπτωση των μοντέλων βαθιάς μάθησης, επιλέγεται συνήθως η χρήση των word embeddings. Κατόπιν, γίνεται η επιλογή του κατάλληλου αλγόριθμου ταξινόμησης (classifier algorithm) για την διαδικασία της εκπαίδευσης του μοντέλου. Γνωστοί αλγόριθμοι ταξινόμησης είναι ο Naive Bayes, η λογιστική παλινδρόμηση (logistic regression), ο Support Vector Machine. Το μοντέλο εκπαιδεύεται σε προ-επισημασμένα δεδομένα και κατόπιν αξιολογείται ως προς την ακρίβεια των προβλέψεων του [78].

## 2.5 Εξαγωγή Πληροφορίας (*Information Extraction*)

Σε μια εποχή με καταιγισμό πληροφοριών η «πληροφορία» μπορεί να σημαίνει πολλά πράγματα, όπως γεγονότα, άτομα ή σχέσεις μεταξύ ανθρώπων, τόπων ή οργανισμών κ.λπ. Η Εξαγωγή πληροφορίας (*Information Extraction - IE*) αναφέρεται στο κομμάτι της NLP που αφορά στην εξαγωγή σχετικών πληροφοριών από έγγραφα κειμένου.

Καθημερινά δεχόμαστε μεγάλη ποσότητα από περιεχόμενο κειμένου, είτε ως συνομιλίες - μηνύματα στο τηλέφωνο, είτε μέσω ηλεκτρονικού ταχυδρομείου ή μέσω μεγαλύτερων κειμένων που διαβάζουμε για διασκέδαση ή στη δουλειά ή για να ενημερωθούμε για τα τρέχοντα υποθέσεων. Τέτοια έγγραφα κειμένου είναι μια πλούσια πηγή πληροφορίας

Σε σύγκριση με δομημένες πηγές πληροφοριών όπως βάσεις δεδομένων ή πίνακες ή ημιδομημένες πηγές όπως π.χ. ιστοσελίδες (που έχουν κάποια σήμανση), το κείμενο είναι μια μορφή μη δομημένων δεδομένων. Για παράδειγμα, σε μια βάση δεδομένων ξέρουμε πού να αναζητήσουμε κάτι με βάση το σχήμα (*scheme*) της. Ωστόσο, σε μεγάλο βαθμό, τα έγγραφα κειμένου περιλαμβάνουν κείμενο ελεύθερης ροής, χωρίς καθορισμένο σχήμα. Αυτό κάνει την εξαγωγή πληροφορίας ένα δύσκολο πρόβλημα. Σε αρκετές περιπτώσεις στις οποίες υπάρχει ένα σταθερό μοτίβο (π.χ. διευθύνσεις, αριθμοί τηλεφώνου, ημερομηνίες κ.λπ.), η εξαγωγή πληροφορίας μπορεί να είναι μια σχετικά απλή διαδικασία, χρησιμοποιώντας τεχνικές εξαγωγής βάσει μοτίβων, όπως κανονικές εκφράσεις (*regular expressions*), ακόμη και αν και το ίδιο το κείμενο θεωρείται μη δομημένα δεδομένα. Ωστόσο, η εξαγωγή άλλων πληροφοριών όπως, π.χ. ονόματα ανθρώπων, σχέσεις μεταξύ διαφορετικών οντοτήτων στο κείμενο, λεπτομέρειες για ένα συμβάν ημερολογίου, κ.λπ. ενδέχεται να απαιτούν πιο προηγμένες λειτουργίες επεξεργασίας γλώσσας.

### 2.5.1 Εφαρμογές Εξαγωγής Πληροφορίας

Η Εξαγωγή Πληροφορίας χρησιμοποιείται σε ένα ευρύ φάσμα εφαρμογών του πραγματικού κόσμου, όπως άρθρα ειδήσεων, κοινωνικά μέσα, επιστημονικές μελέτες κ.α. Μερικές από αυτές είναι:

#### **Προσθήκη ετικετών σε ειδήσεις και άλλο περιεχόμενο (Tagging news and other content)**

Υπάρχει μεγάλη ποσότητα κειμένου που δημιουργείται για διάφορα γεγονότα που συμβαίνουν σε όλο τον κόσμο κάθε μέρα. Εκτός από την ταξινόμηση του κειμένου, είναι χρήσιμο για ορισμένες εφαρμογές, όπως π.χ. μηχανές αναζήτησης, εάν τέτοια κείμενα επισημαίνονται με σημαντικές οντότητες που αναφέρονται σε αυτά. Για παράδειγμα, στην παρακάτω εικόνα από ένα άρθρο σχετικό με τις πρόσφατες πλημμύρες\* εμφανίζονται οι ετικέτες αυτές, ώστε ο αναγνώστης να μπορεί να μεταβεί απευθείας στην συγκεκριμένη οντότητα

**TAGS: Κακοκαιρία Πλημμύρες Βόλος Εύβοια Ελλάδα**

Σχήμα 2: Ετικέτες

(\*<https://www.news247.gr/ellada/kakokairia-pnigontai-volos-evvoia-kai-jthiotida-nixta-agonias-gia-iperxeilisi-potamon/>)

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

## Chatbots

Ένα chatbot πρέπει να κατανοήσει την ερώτηση του χρήστη για να δημιουργήσει/ανακτήσει μια σωστή απάντηση. Για παράδειγμα, στην ερώτηση, «Ποια είναι τα καλύτερα εστιατόρια κοντά στο Σύνταγμα;» το chatbot πρέπει να καταλάβει ότι το "Σύνταγμα" είναι τοποθεσία όπως και τον όρο «εστιατόρια», ώστε στη συνέχεια να προσδιορίσει τα εστιατόρια σε μια ορισμένη απόσταση από Σύνταγμα. Η Εξαγωγή Πληροφορίας είναι χρήσιμη για την αντιστοίχιση/εξαγωγή τέτοιων συγκεκριμένων πληροφοριών από μια ομάδα διαθέσιμων δεδομένων.

## Εφαρμογές στα μέσα κοινωνικής δικτύωσης

Πολλές πληροφορίες διαδίδονται μέσω των καναλιών κοινωνικής δικτύωσης, όπως το Twitter. Η εξαγωγή ενημερωτικών αποσπασμάτων από το κείμενο των μέσων κοινωνικής δικτύωσης μπορεί να βοηθήσει στην ανάλυση των γεγονότων την ώρα που συμβαίνουν, όπως π.χ. σε μια φυσική καταστροφή.

## Εργασίες Εξαγωγής Πληροφορίας

Η Εξαγωγή Πληροφορίας (IE – Information Extraction) είναι ένας όρος που χρησιμοποιείται για να περιγράψει σε μια σειρά διαφορετικών εργασιών διαφορετικής πολυπλοκότητας. Ο πρωταρχικός στόχος της Εξαγωγής Πληροφορίας είναι να εξάγει «γνώση» από το κείμενο και καθεμία από αυτές τις εργασίες παρέχει διαφορετικές πληροφορίες για να επιτευχθεί ο στόχος. Στο παρακάτω απόσπασμα από ενημερωτική ιστοσελίδα \*

Ο Ming-Chi Kuo, αναλυτής της Apple στην TF International Securities, είπε ότι **οποιοδήποτε ζήτημα υπερθέρμανσης οφείλεται πιθανώς σε «συμβιβασμούς» στη σχεδίαση του τηλεφώνου** καθώς η εταιρεία προσπάθησε να μειώσει το βάρος του προϊόντος. Το μοντέλο iPhone Pro έχει ένα πλαίσιο κατασκευασμένο από τιτάνιο, έναν κακό αγωγό θερμότητας, ο οποίος θα μπορούσε να εμποδίσει τη διάχυση της θερμότητας στο ακουστικό. Προηγουμένως, η Apple είχε χρησιμοποιήσει ανοξείδωτο ατσάλι για τα υπερσύγχρονα τηλέφωνα της.

Είπε ότι **η Apple πιθανότατα θα διευθετήσει το πρόβλημα μέσω ενημερώσεων λογισμικού**. Ο Kuo είπε ότι το αναφερόμενο πρόβλημα είναι απίθανο να προκλήθηκε από ένα νέο τσιπ που κατασκευασμένο στην Ταϊβάν που χρησιμοποιήθηκε στα μοντέλα Pro.

Ωστόσο, ο αναλυτής ανέφερε ότι **το πρόβλημα θα μπορούσε να επηρεάσει τις πωλήσεις του iPhone Pro εάν δεν αντιμετωπιστεί**.

Το iPhone 15 ξεκίνησε να πωλείται την περασμένη εβδομάδα. Στο Ηνωμένο Βασίλειο, κοστίζει από 799 £ (924€) για το βασικό μοντέλο έως 999 £ (1.155€) για το Pro και 1.199 £ (1.387€) για το μεγαλύτερο Pro Max. «Εάν η Apple δεν αντιμετωπίσει σωστά αυτό το ζήτημα, θα μπορούσε να επηρεάσει αρνητικά τις αποστολές κατά τη διάρκεια του κύκλου ζωής του προϊόντος της σειράς iPhone 15 Pro», δήλωσε ο Kuo.

Σχήμα 3: Άρθρο

(\*<https://www.news247.gr/technologia/oi-xristes-iphone-15-paraponiountai-gia-iperthermansi-ton-siskevon-ti-apanta-i-apple/>)

ένας αναγνώστης μπορεί να αναγνωρίσει ότι το άρθρο αναφέρεται στην εταιρεία Apple, ότι ο Ming-Ching Kuo είναι άνθρωπος, ότι 799£ είναι ποσό σε λίρες Αγγλίας, όπως και το 924€ είναι ποσό σε Ευρώ.

Για την αντίστοιχη κατανόηση με μηχανικό τρόπο, πρέπει να προσδιοριστεί (εξαχθεί) η πληροφορία ότι Apple -> Εταιρεία, Ming-Ching Kuo -> άτομο & αναλυτής της Apple, 799£ -> χρηματικό ποσό. Για να υλοποιηθεί αυτή εξαγωγή απαιτούνται διαφορετικά επίπεδα ΙΕ.

Ο προσδιορισμός ότι το άρθρο αφορά την «αγορά» ή την «τιμή μετοχών» σχετίζεται με την εργασία ΙΕ εξαγωγή λέξης-κλειδιού ή φράσης-κλειδιού (Keyword or Keyphrase Extraction - KPE). Η αναγνώριση της Apple ως οργανισμού και του Ming-Chi Kuo ως άτομο εμπίπτει στην εργασία ΙΕ της αναγνώρισης ονοματισμένης οντότητας (NER). Αναγνωρίζοντας ότι η λέξη Apple δεν αναφέρεται στο φρούτο, αλλά στην εταιρεία Apple, Inc. είναι δουλειά της αποσαφήνισης και σύνδεσης οντοτήτων (Named entity disambiguation and linking). Εξάγοντας τις πληροφορίες ότι ο Ming-Chi Kuo είναι αναλυτής της Apple που εργάζεται στη TF International Securities Maestri απαιτεί την εργασία ΙΕ της εξόρυξης σχέσεων (relation extraction).

Υπάρχουν κι άλλες εργασίες ΙΕ όπως η εξαγωγή γεγονότων (event extraction), η χρονική εξαγωγή πληροφορίας (temporal information extraction) κ.α.

Κάθε μία από αυτές τις εργασίες απαιτεί διαφορετικά επίπεδα γλωσσικής επεξεργασίας. Μια σειρά από μεθόδους βάσει κανόνων (rule-based) καθώς και εποπτευόμενης, μη εποπτευόμενης και ημι-εποπτευόμενης μηχανικής μάθησης και βαθιάς μάθησης μπορεί να χρησιμοποιηθεί για την ανάπτυξη λύσεων για αυτές τις εργασίες.

Με την ανάπτυξη των LLMs (Large Language Models – Μεγάλα Γλωσσικά Μοντέλα) αρκετές από τις εργασίες της NLP «ενσωματώθηκαν» σε αυτά και εκτελούνται «αόρατα» από τον χρήστη με μεγάλη ταχύτητα και παράλληλη εκτέλεση.

### 2.5.2 *Keyphrase extraction (εξαγωγή φράσεων-κλειδιών)*

Η εξαγωγή φράσεων ή λέξεων κλειδιών (keyphrase ή keyword extraction) είναι μια διαδικασία που περιλαμβάνει την αυτόματη αναγνώριση και εξαγωγή των πιο σχετικών φράσεων από ένα έγγραφο, οι οποίες συνοψίζουν το περιεχόμενό του [82]. Αυτές οι φράσεις, γνωστές ως φράσεις κλειδιά, παρέχουν μια συμπαγή παρουσίαση των κύριων θεμάτων του εγγράφου και είναι ιδιαίτερα χρήσιμες σε διάφορες εφαρμογές σε συστήματα διαχείρισης ψηφιακών πληροφοριών. Οι φράσεις κλειδιά μπορούν να χρησιμοποιηθούν για τη δημιουργία ευρετηρίου μιας συλλογής εγγράφων βάσει περιεχομένου, επιτρέποντας σημασιολογική και πολύπλευρη αναζήτηση. Είναι σημαντικές για τη βελτίωση της προσβασιμότητας και της οργάνωσης μεγάλων συλλογών κειμένων, παρέχοντας ετικέτες (tags) που συνοψίζουν και κατηγοριοποιούν αποτελεσματικά το περιεχόμενο. Διευκολύνουν έτσι την ομαδοποίηση παρόμοιων εγγράφων και τις γρήγορες περιλήψεις.

Για την υλοποίηση αυτής της σημαντικής, για την NLP, διαδικασίας υπάρχουν τρεις κατηγορίες μεθόδων:

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

*Επιβλεπόμενες μέθοδοι (Supervised Methods):* Αυτές βασίζονται σε προ-επισημασμένα (on pre-labeled) σύνολα δεδομένων για την εκπαίδευση μοντέλων που μπορούν να προβλέψουν τις φράσεις-κλειδιά.

*Μη-επιβλεπόμενες μέθοδοι (Unsupervised Methods):* Δεν χρησιμοποιούν προ-επισημασμένα δεδομένα και μπορεί να κάνουν χρήση στατιστικών εργαλείων, όπως TF-IDF μέθοδος που αναφέρθηκε νωρίτερα ή μεθόδους βασισμένους σε γράφους, ώστε να εκμεταλλευθούν τις σχέσεις μεταξύ των λέξεων για την αναγνώριση των φράσεων-κλειδιών.

Εφαρμόζονται, επίσης, *προσεγγίσεις βαθιάς μάθησης*, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα, τα οποία μπορούν να συλλάβουν το πλαίσιο και τη σειρά των λέξεων πιο αποτελεσματικά και οι οποίες χρησιμοποιούνται από τα LLMs για σειρά εργασίες.

### 2.5.3 Name Entity Recognition (NER)

Όπως αναφέρθηκε και νωρίτερα, η αναγνώριση ονοματισμένης οντότητας – Name Entity Recognition (NER) είναι η εργασία της NLP που αναγνωρίζει τις οντότητες σε ένα κείμενο. Υπό τον όρο οντότητες εντοπίζουμε, συνήθως, ονόματα ατόμων, τοποθεσιών (χώρες, πόλεις π.χ.), οργανισμών, νομισμάτων, χρονικών εκφράσεων κ.α. Είναι βασικό κομμάτι στην αλυσίδα των εφαρμογών που βασίζονται στην NLP και ειδικά στην εργασία εξαγωγή πληροφορίας.

Για την υλοποίηση της NER διάφορες προσεγγίσεις έχουν χρησιμοποιηθεί [84], όπως εξελίσσεται και αναπτύσσεται η NLP. Οι παραδοσιακές μέθοδοι περιλαμβάνουν:

*Rule-based Approaches (προσεγγίσεις βάσει κανόνων):* Στα συστήματα αυτά κανόνες ορισμένοι χειροκίνητα χρησιμοποιούνται για να αναγνωριστούν οντότητες, βάσει μοτίβων και αναζητήσεων σε λεξικά. Π.χ. με την χρήση Regular Expressions (κανονικών εκφράσεων) για αναγνώριση μοτίβων.

*Unsupervised Learning Approaches (προσεγγίσεις μη-επιβλεπόμενης μάθησης):* Εδώ χρησιμοποιούνται μη επισημασμένα δεδομένα για εκπαίδευση και συνήθως γίνεται χρήση ομαδοποίησης (clustering) ή παρόμοιων αλγορίθμων για αναγνώριση παρόμοιων οντοτήτων.

*Feature-based Supervised Learning Approaches (επιβλεπόμενες προσεγγίσεις βασισμένες σε χαρακτηριστικά):* Στην περίπτωση αυτή, η NER αντιμετωπίζεται είτε ως ταξινόμηση (classifying) κάθε τμήματος κειμένου σε διαφορετικές κατηγορίες, είτε ως επισήμανση (labelling) κάθε μέρους του κειμένου ακολουθιακά. Λαμβάνοντας υπόψη τα επισημασμένα δείγματα δεδομένων, τα χαρακτηριστικά (features) έχουν σχεδιαστεί προσεκτικά για να αντιπροσωπεύουν κάθε εκπαιδευτικό παράδειγμα. Κατόπιν, αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για την εκπαίδευση του μοντέλου για την αναγνώριση παρόμοιων μοτίβων σε άγνωστα δεδομένα.

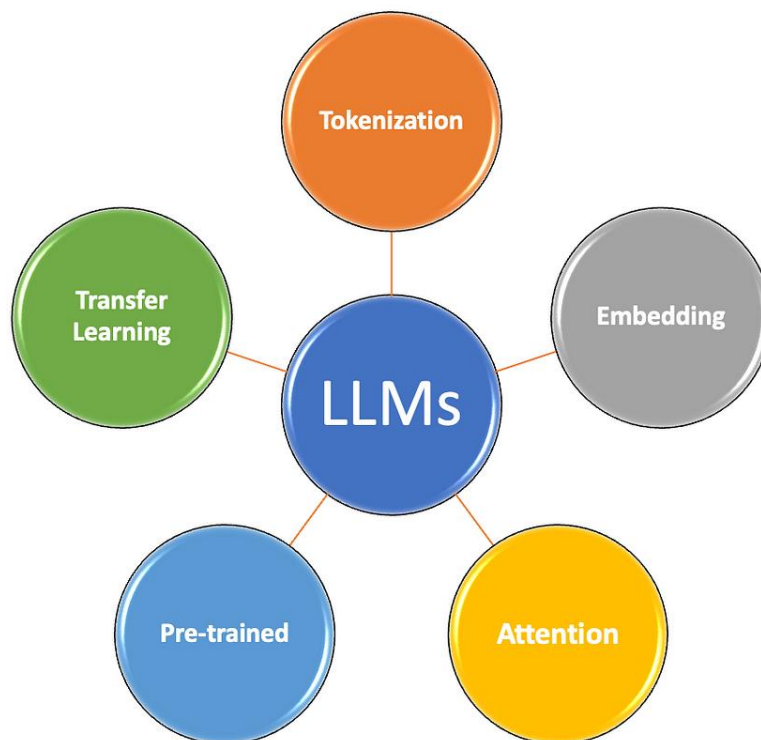
*Deep Learning Approaches (προσεγγίσεις βαθιάς μάθησης):* οι εξελίξεις στη βαθιά μάθηση έδωσαν μεγάλη ώθηση στην ανάπτυξη της NER. Τα μοντέλα που βασίζονται σε Deep Learning είναι καταρχάς, σε θέση να μαθαίνουν πολύπλοκα χαρακτηριστικά από δεδομένα μέσω μη γραμμικών συναρτήσεων ενεργοποίησης και κατά δεύτερον, είναι αποτελεσματικά στην αυτόματη μάθηση χρήσιμων αναπαραστάσεων από τα δεδομένα. Έτσι, δεν απαιτείται πλέον ο σχεδιασμός των χαρακτηριστικών, μια διαδικασία απαιτητική και χρονοβόρα, όπως στις παραδοσιακές μεθόδους.

## 2.6 LLMs (*Large Language Models – Μεγάλα Γλωσσικά Μοντέλα*)

Τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) είναι προηγμένα συστήματα τεχνητής νοημοσύνης, σχεδιασμένα ειδικά για να κατανοούν, να παράγουν και να αλληλοεπιδρούν με την ανθρώπινη γλώσσα. Τα μοντέλα αυτά χαρακτηρίζονται «μεγάλα» τόσο λόγω μεγέθους των νευρωνικών δικτύων πάνω στα οποία βασίζονται (το GPT-3 έχει 175 δισεκατομμύρια παραμέτρους) όσο και των τεράστιων ποσοτήτων δεδομένων στα οποία εκπαιδεύονται. Ενδεικτικά, το GPT-3 υπολογίζεται ότι έχει εκπαιδευτεί σε περίπου 45TB από διάφορες πηγές ή αλλιώς σε  $\approx 499$  δισεκατομμύρια tokens [15, 76].

Τα LLMs βασίζονται στη βαθιά μάθηση (Deep Learning) έναν κλάδο της μηχανικής μάθησης που χρησιμοποιεί νευρωνικά δίκτυα, τα οποία είναι μαθηματικά μοντέλα εμπνευσμένα από τη δομή και τη λειτουργία του εγκεφάλου, στην επεξεργασία δεδομένων και στη δημιουργία μοτίβων για τη λήψη αποφάσεων.

Τα Μεγάλα Γλωσσικά Μοντέλα εκπαιδεύονται για την εκτέλεση ποικίλων εργασιών, όπως η σύνταξη κώδικα, η δημιουργία περίληψης κειμένου κ.α. Τα LLMs προ-εκπαιδεύονται (pre-trained) και στη συνέχεια τελειοποιούνται (fine-tuned), ώστε να μπορούν να λύσουν προβλήματα ταξινόμησης κειμένου, απάντησης ερωτήσεων, σύνοψης εγγράφων και δημιουργίας κειμένου [65]. Οι δυνατότητές τους μπορούν να χρησιμοποιηθούν σε πληθώρα εφαρμογών NLP όπως μετάφρασης, chatbots, βοηθών AI κ.α. [69]



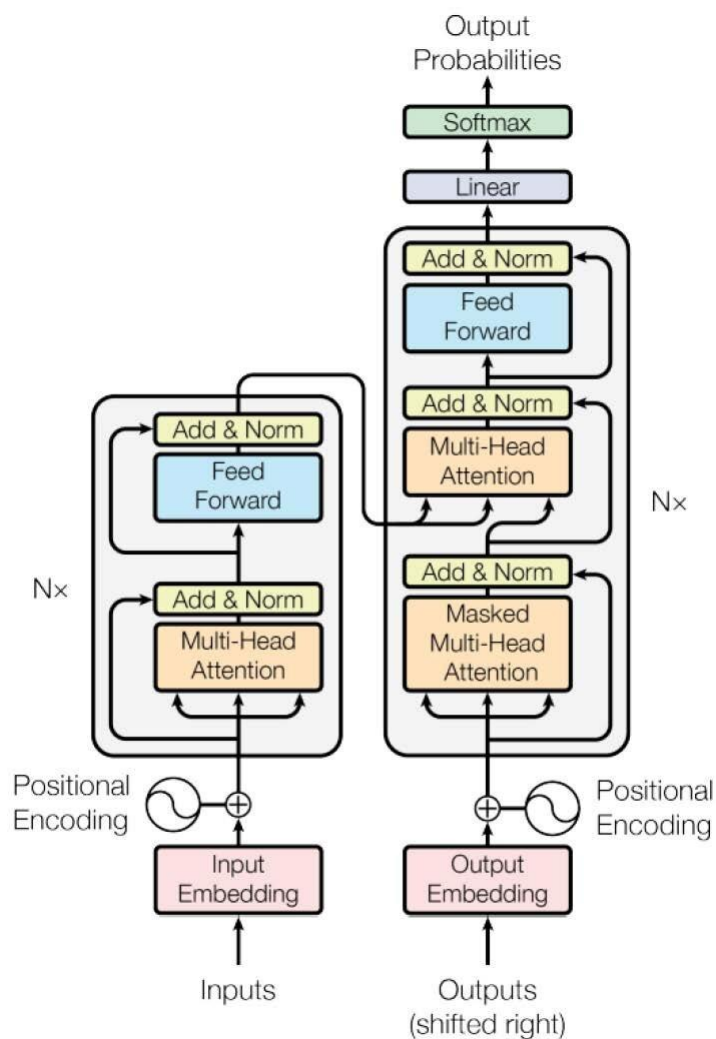
Σχήμα 4: Βασικές συνιστώσες των LLMs [σχήμα από 6]



### 2.6.1 Αρχιτεκτονική Transformer

Το μεγάλο άλμα στην εξέλιξη των Μεγάλων Γλωσσικών Μοντέλων έγινε με την εισαγωγή της αρχιτεκτονικής Transformer (μετασχηματιστή), όπως αυτή παρουσιάστηκε το 2017 στο άρθρο «Attention Is All You Need» των Vaswani κ.α.[5] και η οποία αποτελεί την βάση για τα σύγχρονα LLMs.

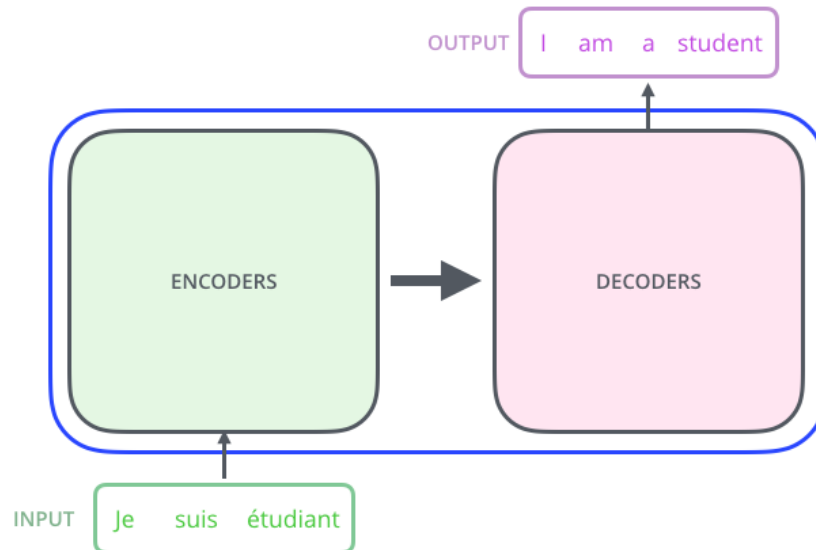
Τα μοντέλα που είναι βασισμένα στην αρχιτεκτονική Transformers έχουν σημαντικά πλεονεκτήματα έναντι των προηγούμενων μοντέλων που στηρίζονταν στην διαδοχική επεξεργασία δεδομένων, όπως τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs) και τα Δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTMs). Στα προηγούμενα μοντέλα η επεξεργασία των δεδομένων γινόταν σειριακά και επιπλέον υπήρχαν περιορισμοί στην αποτύπωση των συνδέσεων μεταξύ των λέξεων και του γενικότερου πλαισίου (context). Αντίθετα, η αρχιτεκτονική Transformers επιτρέπει ολόκληρες ακολουθίες να επεξεργάζονται ταυτόχρονα, βελτιώνοντας σημαντικά την αποδοτικότητα και την αποτελεσματικότητα, ιδιαίτερα για μακρύτερες ακολουθίες. Ο μηχανισμός δε αυτό-προσοχής (self-attention) που θα αναλυθεί παρακάτω βοηθά στην πολύ πιο αποτελεσματική αποτύπωση του γενικού πλαισίου και σε βελτιωμένη απόδοση σε εργασίες επεξεργασίας φυσικής γλώσσας



Σχήμα 5: Αρχιτεκτονική Transformer [σχήμα από 5]

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Η αρχιτεκτονική Transformer αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή, όπως φαίνεται στο παρακάτω σχήμα:



Σχήμα 6: Κωδικοποιητές - Αποκωδικοποιητές Transformer [σχήμα από 4]

Ένα μοντέλο μετασχηματιστή (Transformer) επεξεργάζεται κειμενικά δεδομένα, αρχικά σπάζοντας σε λεκτικές μονάδες (tokens) την είσοδο και στη συνέχεια, διεξάγοντας ταυτόχρονα μαθηματικές εξισώσεις για να ανακαλύψει σχέσεις μεταξύ των tokens. Έτσι, το μοντέλο δεν εξάγει απλά την πιο συχνή επόμενη λέξη (όπως είναι π.χ. οι προτάσεις από το πληκτρολόγιο των κινητών τηλεφώνων), αλλά διατηρεί την έννοια του πλαισίου συζήτησης (context), ώστε το παραγόμενο κείμενο να είναι συναφές και συνεκτικό.

Το πρώτο βήμα για την υλοποίηση του είναι το tokenization. Tokenization είναι η διαίρεση του κειμένου σε μικρότερες μονάδες που ονομάζονται tokens. Τα tokens περιλαμβάνουν όλες τις λέξεις, τα σημεία στίξης, κ.λπ. Το βήμα του tokenization παίρνει κάθε λέξη, πρόθεμα, επίθημα και σημείο στίξης και το αντιστοιχίζει σε ένα token από τη βιβλιοθήκη. Στόχος του είναι να δημιουργήσει μια τυποποιημένη αναπαράσταση για αποτελεσματική επεξεργασία και ανάλυση [7].

## Tokenization

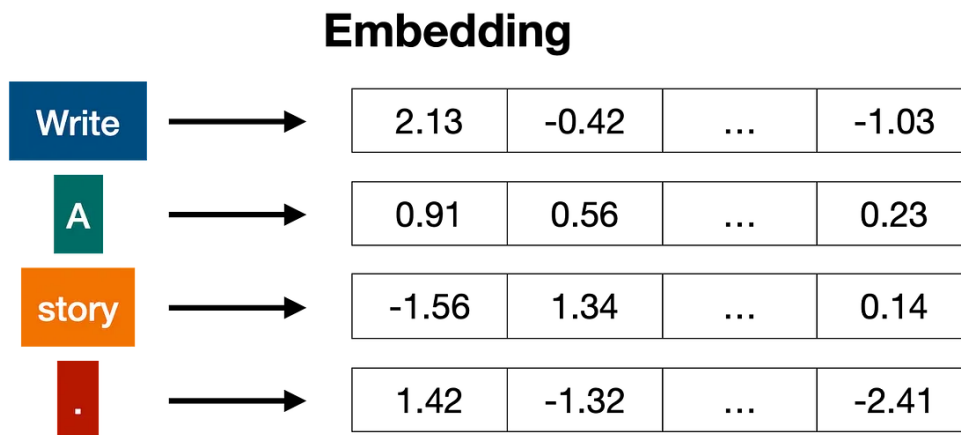


Σχήμα 7: Tokenization [σχήμα από 81]

Το επόμενο στάδιο περιλαμβάνει την μετατροπή των tokens σε μια μορφή αριθμητικής αναπαράστασης (συνήθως σε διανύσματα), ώστε να είναι δυνατή η χρήση και επεξεργασία τους



από τα νευρωνικά δίκτυα. Τα tokens μετατρέπονται σε διανύσματα (vectors) με την χρήση ενός αλγόριθμου embedding (ενσωμάτωσης), όπως π.χ. το word2vec. Το word embedding είναι ένας τρόπος μετατροπής των tokens σε διανύσματα. Ο στόχος είναι η καταγραφή σημασιολογικών σχέσεων και πληροφοριών με βάση τα συμφραζόμενα. Αντιπροσωπεύουν διαφορετικούς τύπους δεδομένων ως σημεία σε έναν πολυδιάστατο χώρο, όπου παρόμοια σημεία δεδομένων συγκεντρώνονται πιο κοντά μεταξύ τους. Αυτές οι αριθμητικές αναπαραστάσεις βοηθούν τις μηχανές να κατανοήσουν και να επεξεργαστούν τα δεδομένα πιο αποτελεσματικά.

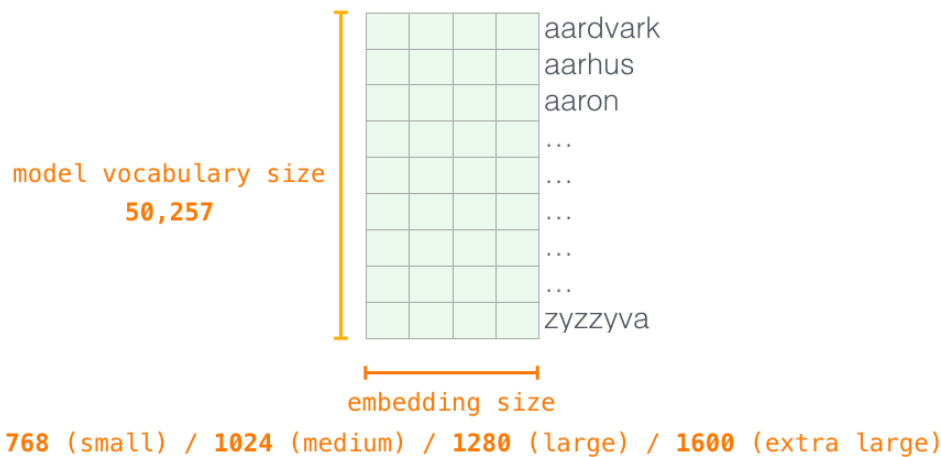


Σχήμα 8: Embeddings [σχήμα από 81]

Ο πρωταρχικός σκοπός των word embeddings στα μοντέλα Transformer είναι να μετατρέψουν κάθε λέξη ή token σε ένα διάνυσμα σταθερού μεγέθους σε έναν χώρο υψηλών διαστάσεων. Αυτή η διανυσματική αναπαράσταση περιλαμβάνει σημασιολογικές πληροφορίες για τη λέξη. Το επίπεδο embedding είναι ουσιαστικά ένας πίνακας αναζήτησης, ο οποίος αντιστοιχίζει λέξεις (ή ακριβέστερα, τις συμβολικές αναπαραστάσεις τους) σε διανύσματα.

Σε αντίθεση με τα επαναλαμβανόμενα νευρωνικά δίκτυα (recurrent neural networks), οι transformers δεν επεξεργάζονται δεδομένα διαδοχικά. Αυτό σημαίνει ότι δεν καταγράφουν εγγενώς τη θέση ή τη σειρά των λέξεων σε μια πρόταση. Για να αντιμετωπιστεί αυτό, οι Transformers προσθέτουν κωδικοποιήσεις θέσης στα word embeddings (ενσωματώσεις λέξεων). Οι κωδικοποιήσεις θέσης (positional encoding) είναι διανύσματα που κωδικοποιούν τη θέση κάθε λέξης στην ακολουθία. Βοηθούν το μοντέλο να κατανοήσει τη σειρά των λέξεων και τις σχετικές θέσεις των λέξεων. Κάθε θέση στην πρόταση έχει μια μοναδική κωδικοποίηση θέσης και αυτή προστίθεται στο word embedding, διασφαλίζοντας ότι το μοντέλο μπορεί να διακρίνει την ίδια λέξη σε διαφορετικές θέσεις. Οι τελικές ενσωματώσεις εισόδου για τον μετασχηματιστή είναι το άθροισμα των word embeddings και των positional encodings. Αυτός ο συνδυασμός παρέχει στο μοντέλο πληροφορίες τόσο για τη σημασία των λέξεων, όσο και για τις θέσεις τους στην πρόταση. Τα embeddings στα μοντέλα αρχιτεκτονικής transformer δεν είναι στατικές αναπαραστάσεις που έχουν προϋπολογιστεί. Κατά τη διάρκεια της εκπαίδευσης των μοντέλων υπολογίζονται αρχικά και, κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, καθώς το μοντέλο μαθαίνει, εξελίσσει / προσαρμόζει τα embeddings για τη βελτιστοποίηση της απόδοσής του. Όταν ο χρήστης χρησιμοποιεί το μοντέλο, τα embeddings προσαρμόζονται εκ νέου, πριν απαντήσει το μοντέλο, ώστε να ανταποκρίνονται καλύτερα στο συγκεκριμένο κείμενο που δόθηκε. Για αυτό και οι διανυσματικές αναπαραστάσεις στα LLMs είναι δυναμικές και όχι στατικές.

## Token Embeddings (wte)



Σχήμα 9: Token Embeddings [σχήμα από 8]

Ένα από τα πιο γνωστά παραδείγματα LLM, το GPT-3 έχει embeddings dimension (διάσταση των ενσωματώσεων) και πλάτος των κρυφών στρωμάτων 12.288. Αυτό σημαίνει ότι κάθε token (λέξη ή μέρος μιας λέξης) αντιπροσωπεύεται από ένα διάνυσμα 12.288 αριθμών. Το αντίστοιχο νούμερο για την μεγαλύτερη εκδοχή του προηγούμενου μοντέλου GPT- 2 είναι 1.600.

### 2.6.1.1 Μηχανισμός Αυτό-προσοχής (Self-Attention)

Τα μοντέλα μετασηματιστών λειτουργούν με μηχανισμούς αυτό-προσοχής (self-attention), γεγονός που τους επιτρέπει να μαθαίνουν πιο γρήγορα από τα παραδοσιακά μοντέλα, όπως τα μοντέλα μακράς βραχυπρόθεσμης μνήμης (LSTMs). Η αυτό-προσοχή είναι η διαδικασία που επιτρέπει στο μοντέλο Transformer να εξετάσει διαφορετικά μέρη της ακολουθίας ή ολόκληρο το πλαίσιο μιας πρότασης για να δημιουργήσει προβλέψεις [9].

Ο σκοπός του επιπέδου self-attention είναι να επιτρέψει στο μοντέλο να σταθμίσει τη σημασία διαφορετικών λέξεων (ή σωστότερα tokens) σε μια πρόταση για μια δεδομένη εργασία, καθιστώντας το ικανό να κοιτάζει άλλες θέσεις στην ακολουθία εισόδου για να κατανοήσει καλύτερα μια λέξη στο γενικό πλαίσιο. Για παράδειγμα, στη φράση «Το βιβλίο που μου χάρισε ο Γιώργος, και που **το** αγόρασε από το βιβλιοπωλείο της γειτονιάς του», ο μηχανισμός αυτό-προσοχής είναι που επιτρέπει την συσχέτιση του δεύτερου «το» στη φράση «που το αγόρασε» με τη λέξη βιβλίο.

Για την υλοποίηση του μηχανισμού Self-Attention, κάθε λέξη στην ακολουθία εισόδου μετασηματίζεται σε τρία διανύσματα: ένα διάνυσμα ερωτήματος - Query vector (Q), ένα διάνυσμα κλειδιού - Key vector (K) και ένα διάνυσμα τιμής - Value vector (V), χρησιμοποιώντας διαφορετικούς γραμμικούς μετασηματισμούς. Για κάθε λέξη προσδιορίζεται η ομοιότητά της (similarity) με κάθε άλλη λέξη στην πρόταση, περιλαμβάνοντας και τον εαυτό της. Η ομοιότητα προσδιορίζεται ως το εσωτερικό γινόμενο (dot product) του Query της λέξης με το Key κάθε άλλης λέξης. Τα εσωτερικά γινόμενα μειώνονται κατά την τετραγωνική ρίζα της διάστασης των βασικών διανυσμάτων για την αποφυγή εξαιρετικά μεγάλων τιμών, οι οποίες μπορεί να εμποδίσουν την εκπαίδευση του μοντέλου. Το αποτέλεσμα που προκύπτει ονομάζεται βαθμολογία

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

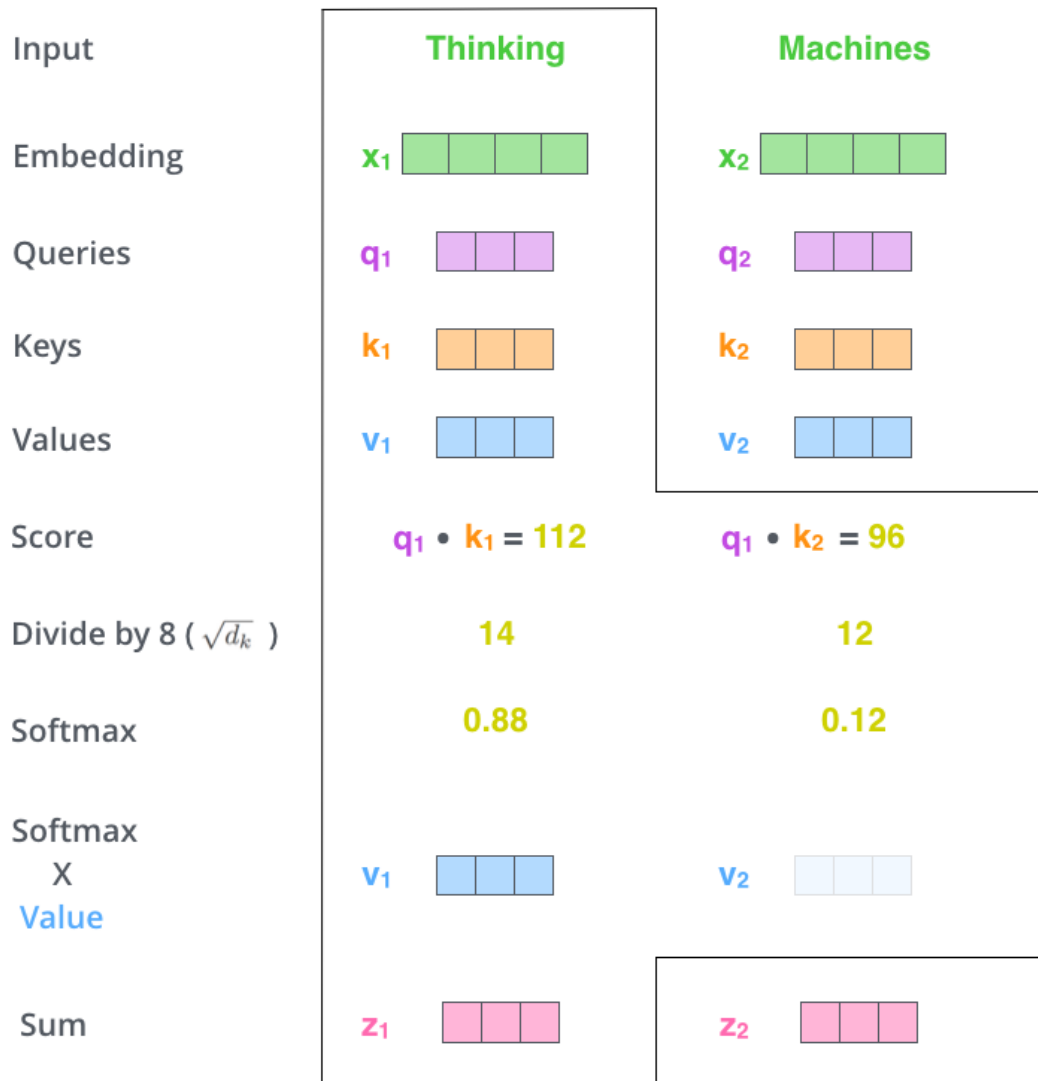
προσοχής (attention score) και υποδεικνύει πόση σημασία πρέπει να δοθεί σε άλλα μέρη της πρότασης εισαγωγής κατά την επεξεργασία μιας συγκεκριμένης λέξης.

Τέλος, εφαρμόζεται η συνάρτηση softmax σε αυτές τις βαθμολογίες για την κανονικοποίηση τους σε πιθανότητες, διασφαλίζοντας ότι το άθροισμα είναι 1.

Η έξοδος του επιπέδου αυτό-προσοχής για κάθε λέξη είναι ένα σταθμισμένο άθροισμα των διανυσμάτων τιμών (Values), όπου τα βάρη είναι οι βαθμολογίες προσοχής κανονικοποιημένες με softmax. Στην έξοδο αυτή «συμμετέχουν» όλες οι υπόλοιπες λέξεις (με τα κατάλληλα βάρη, όπως αναφέρθηκε) δίνοντας έτσι το πλαίσιο (context) στη λέξη.

Καθώς τα βάρη για τον υπολογισμό των τιμών Query, Key και Value είναι ίδια για όλα τα tokens, ο υπολογισμός τους μπορεί να γίνεται παράλληλα και άρα πιο γρήγορα.

Η έξοδος του επιπέδου self-attention προστίθεται στο άθροισμα των word embeddings και των positional encodings, το οποίο και υπολογίστηκε σε προηγούμενο στάδιο. Το αποτέλεσμα που προκύπτει ονομάζεται residual connection και αυτό χρησιμοποιείται στα επόμενα στάδια, αφού περιέχει όλες τις πληροφορίες που χρειάζονται.



Σχήμα 10: Self-Attention Softmax [σχήμα από 4]

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Συνοπτικά, ο μηχανισμός attention μετακινεί τις λέξεις (την διανυσματική τους αναπαράσταση) σε μια πρόταση (ή κομμάτι κειμένου) πιο κοντά, όταν αυτές σχετίζονται. Με αυτόν τον τρόπο, η λέξη «μάρκα» στην πρόταση «Κέρδισα μια μάρκα στο καζίνο» θα μετακινηθεί πιο κοντά στην λέξη «καζίνο». Αντίστοιχα, στην πρόταση «Η μάρκα της φούστας είναι άγνωστη», η λέξη «μάρκα» θα πλησιάσει τη λέξη «φούστας». Η τροποποιημένη λέξη "μάρκα" σε καθεμία από τις δύο προτάσεις θα φέρει ορισμένες από τις πληροφορίες των γειτονικών λέξεων, προσθέτοντας το πλαίσιο (τα συμφραζόμενα) σε αυτήν. Ο μηχανισμός attention που χρησιμοποιείται στα μοντέλα transformers είναι στην πραγματικότητα πολύ πιο ισχυρός από αυτόν που περιεγράφηκε και ονομάζεται multi-head attention. Στο multi-head attention, πολλά διαφορετικά επίπεδα self-attention, με τα δικά τους σεντ από βάρη, χρησιμοποιούνται για να συλλάβουν τις διαφορετικές σχέσεις μεταξύ των λέξεων.. Ο μηχανισμός αυτός έχει βοηθήσει τα μεγάλα γλωσσικά μοντέλα να φτάσουν σε πολύ υψηλότερα επίπεδα αποτελεσματικότητας κατά την επεξεργασία και τη δημιουργία κειμένου.

#### 2.6.1.2 Encoder - Decoder (Κωδικοποιητής – Αποκωδικοποιητής)

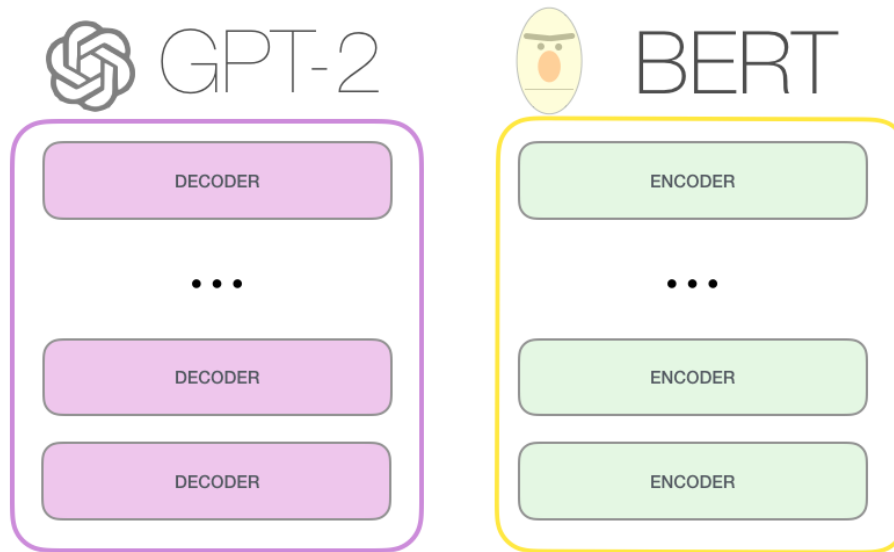
Όπως είδαμε στα Σχήματα [Σχήμα 5] και [Σχήμα 6], το κομμάτι του κωδικοποιητή και του αποκωδικοποιητή έχουν παρόμοια αρχιτεκτονική. Και στον αποκωδικοποιητή οι πιθανές έξοδοι περνούν από τη διαδικασία των Word Embeddings->Positional Embeddings-> Self-Attention->Residual Connections (με διαφορετικά βάρη από τα αντίστοιχα του κωδικοποιητή).

Πρέπει όμως να μεταφερθεί στον αποκωδικοποιητή η πληροφορία για το ποια tokens από την είσοδο είναι τα πλέον σημαντικά. Για τον λόγο αυτό χρησιμοποιείται ο μηχανισμός encoder – decoder attention. Ουσιαστικά, είναι μια ακόμα χρήση του γνωστού μηχανισμού attention, με Query στην έξοδο και Key values στα tokens εισόδου (χρησιμοποιώντας ένα ακόμα σεντ από διαφορετικά βάρη). Η διαδικασία καταλήγει σε νέο σεντ από residual connections τα οποία περιέχουν πληροφορία και για την σχέση εισόδου με έξοδο. Η πληροφορία αυτή είναι εξαιρετικά σημαντική για τον καθορισμό της πιθανής εξόδου.

Τα residual connections οδηγούνται σε Feed Forward Network και κατόπιν, μέσω της χρήσης της Softmax συνάρτησης υπολογίζεται η πιο πιθανή έξοδος.

#### 2.6.1.3 Encoder-only models και Decoder-only models

Από την αρχική αρχιτεκτονική του μετασχηματιστή (Transformer) διαπιστώθηκε ότι είναι δυνατόν να λειτουργήσουν μοντέλα έχοντας είτε μόνο το κομμάτι του κωδικοποιητή (encoder), είτε μόνο το κομμάτι του αποκωδικοποιητή (decoder). Δύο από τα πιο γνωστά μοντέλα που βασίζονται σε αυτή την αρχιτεκτονική είναι το BERT [13] και η οικογένεια GPT (GPT-2 [14], GPT-3 [15] κ.λπ.).



Σχήμα 11: GPT-2 , BERT Αρχιτεκτονική Transformer [σχήμα από 12]

Το BERT χρησιμοποιεί μόνο το κομμάτι του κωδικοποιητή από την αρχιτεκτονική Transformer, γι' αυτό και η έξοδος του είναι σε embeddings και όχι κείμενο. Συνεπώς, δεν μπορεί να χρησιμοποιηθεί για εργασίες όπως παραγωγή κειμένου ή μετάφραση. Είναι δυνατόν όμως η έξοδος του να χρησιμοποιηθεί με τεχνικές, όπως η ομοιότητα συνημίτονου (cosine similarity), για τη σύγκριση μεταξύ embeddings και τον υπολογισμό βαθμού ομοιότητας (similarity score).

Κατά τη διάρκεια της προ-εκπαίδευσής του, το BERT χρησιμοποιεί δύο βασικές διεργασίες:

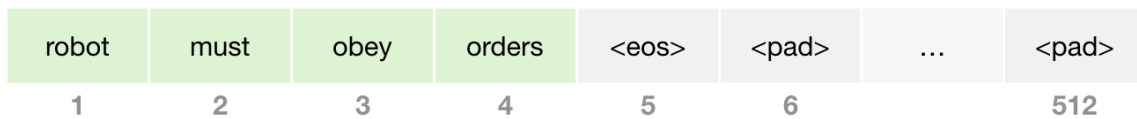
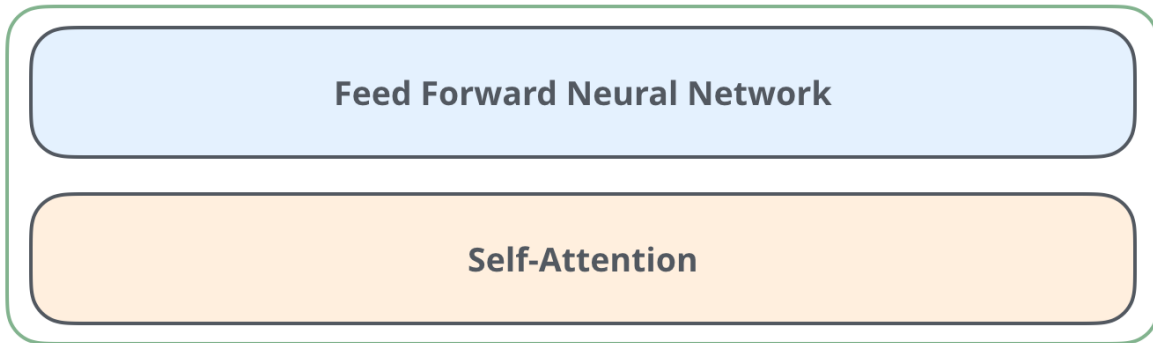
- *Μοντέλο μάσκας γλώσσας (MLM - Masked Language Model)*: Εδώ, τυχαίες λέξεις σε μια πρόταση είναι καλυμμένες – masked (κρυμμένες) και το μοντέλο προσπαθεί να τις προβλέψει με βάση το πλαίσιο που παρέχεται από τις άλλες, μη καλυμμένες λέξεις της πρότασης. Αυτό ωθεί το μοντέλο να αναπτύξει μια βαθιά κατανόηση του πλαισίου και των σχέσεων μεταξύ των λέξεων.
- *Πρόβλεψη επόμενης πρότασης - Next Sentence Prediction (NSP)*: Το μοντέλο έχει δύο προτάσεις και πρέπει να προβλέψει εάν η δεύτερη πρόταση είναι η λογική συνέχεια της πρώτης. Αυτό βοηθά το BERT να κατανοήσει τη σχέση μεταξύ των προτάσεων και βελτιώνει κατά πολύ την απόδοσή του.

Για αυτές τις εργασίες, το BERT ξεκινά με ενσωματώσεις που συνδυάζουν τόσο τις ενσωματώσεις λέξεων όσο και τις ενσωματώσεις θέσης. Οι ενσωματώσεις λέξεων επιτρέπουν στο μοντέλο να κατανοήσει τη σημασία μεμονωμένων λέξεων, ενώ οι ενσωματώσεις θέσης δίνουν στο μοντέλο πληροφορίες σχετικά με τη θέση κάθε λέξης σε μια πρόταση, η οποία είναι κρίσιμη για την κατανόηση της αλληλουχίας και της δομής της γλώσσας.

Αυτές οι ενσωματώσεις δεν "υπολογίζονται εκ των προτέρων" και στη συνέχεια διορθώνονται. Αντίθετα, αποτελούν μέρος των παραμέτρων του μοντέλου και ρυθμίζονται με ακρίβεια κατά τη διαδικασία της προ-εκπαίδευσης. Η προσέγγιση αυτή επιτρέπει στο BERT να δημιουργεί ενσωματώσεις με ευαισθησία στο πλαίσιο για λέξεις που βασίζονται στο περιβάλλον κείμενο, οδηγώντας στην ισχυρή απόδοσή του σε ένα ευρύ φάσμα εργασιών NLP.

## THE TRANSFORMER

### ENCODER BLOCK

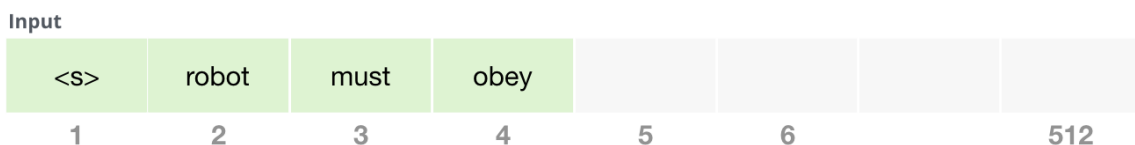
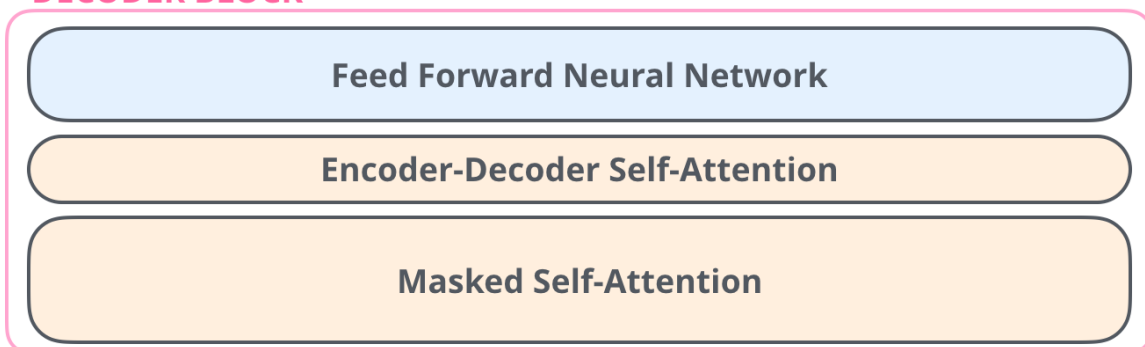


Σχήμα 12: Encoder Block – Transformer [σχήμα από 4]

Αντίθετα, τα μοντέλα της οικογένειας GPT (Generative Pre-Trained Transformer) χρησιμοποιούν το κομμάτι του αποκωδικοποιητή από την αρχιτεκτονική Transformer και μπορούν να παράξουν συνεκτικό και σχετικό με τα συμφραζόμενα κείμενο. Για αυτό, είναι ιδανικά για χρήσης σε εφαρμογές συνομιλίας και δημιουργίας κειμένου.

## THE TRANSFORMER

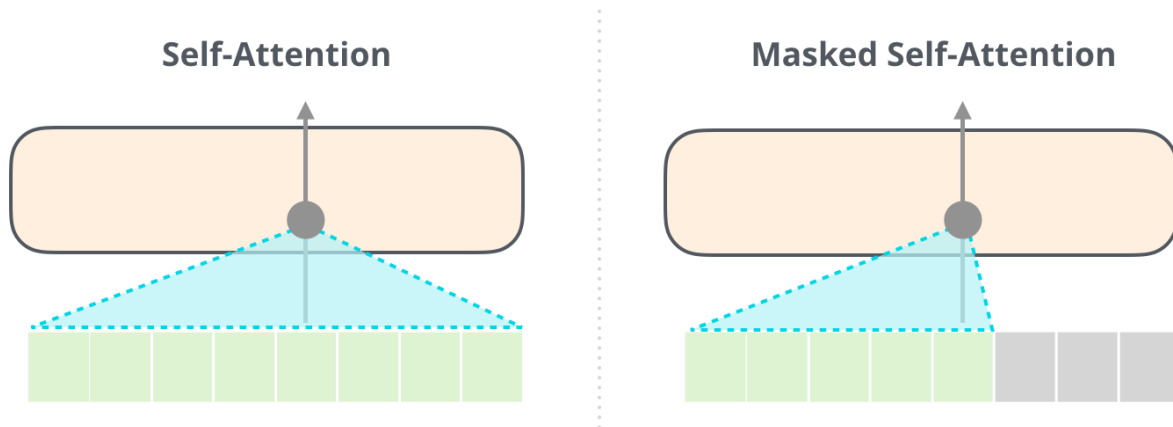
### DECODER BLOCK



Σχήμα 13: Decoder Block – Transformer [σχήμα από 4]

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Από πλευράς αρχιτεκτονικής, μια επιπλέον διαφορά είναι στο είδος του επιπέδου attention που χρησιμοποιούν. Το BERT χρησιμοποιεί το self-attention και το GPT το masked self-attention. Ένα μπλοκ self-attention επιτρέπει σε ένα token να δει τα tokens στα δεξιά της. Στην περίπτωση του masked self-attention αυτό δεν συμβαίνει και έχει πρόσβαση μόνο στα tokens που έχουν προηγηθεί. Για λόγους σύγκρισης, αναφέρουμε ότι το μεγαλύτερο μοντέλο BERT έχει 24 επίπεδα attention, το GPT-2 έχει 12 επίπεδα attention, ενώ το GPT-3 96 επίπεδα.



Σχήμα 14: Self-Attention and Masked Self-Attention [σχήμα από 12]

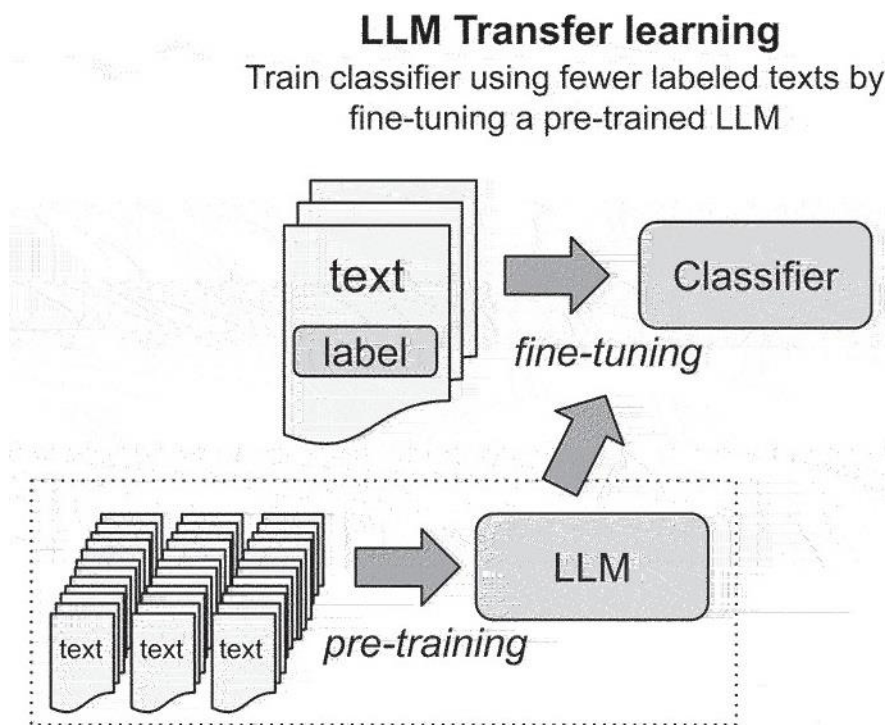
#### 2.6.1.4 Transfer Learning (Μεταφορά Γνώσης)

Τα Μεγάλα Γλωσσικά Μοντέλα προ-εκπαιδεύονται (pre-trained). Η διαδικασία αυτή περιλαμβάνει την εκπαίδευση ενός μοντέλου σε ένα μεγάλο σύνολο κειμένου χωρίς επισήμειωση (unlabeled text) για την εκμάθηση γενικών γλωσσικών αναπαραστάσεων. Ο στόχος είναι να συλλάβει τις στατιστικές ιδιότητες της γλώσσας και να τις κωδικοποιήσει στις παραμέτρους του μοντέλου. Η προ-εκπαίδευση εφαρμόζεται ως βάση για εργασίες όπως η δημιουργία κειμένου, η ανάλυση συναισθημάτων και η απάντηση ερωτήσεων, επιτρέποντας καλύτερη μεταφορά μάθησης και βελτιωμένη απόδοση.

Βασιζόμενοι στην υπάρχουσα πρότερη γνώση είναι δυνατόν να χτιστούν πιο ακριβείς ταξινομητές (classifiers). Λόγω του ρόλου των LLMs ως βάση (foundation) για την προσαρμογή τους σε διάφορες εργασίες, αναφέρονται επίσης και ως foundation models (μοντέλα βάσης). [10]

Το παρακάτω σχήμα αναπαριστά τη χρήση των LLMs στην μεταφορά γνώσης.





Σχήμα 15: LLM Transfer Learning [σχήμα από 10]

Το μοντέλο, έχοντας ήδη προ-εκπαιδευτεί, μπορεί σε επόμενο βήμα να *τελειοποιηθεί* (fine-tuned), ώστε να εκτελεί συγκεκριμένες εργασίες (πχ. ταξινόμησης). Η διαδικασία finetuning γίνεται με προσημειωμένα (Labeled) παραδείγματα συνήθως, εκμεταλλευόμενη ήδη αποκτηθείσα γενική γνώση του μοντέλου. Έτσι, το μέγεθος των δεδομένων που απαιτούνται για το finetuning είναι σημαντικά μικρότερο από την αρχική εκπαίδευση του μοντέλου. Για παράδειγμα, το GPT-3 έχει εκπαιδευθεί σε 45TB δεδομένων από διάφορες πηγές. Για το fine tuning η OpenAI προτείνει να ξεκινήσει από 50 παραδείγματα. [11]

## 2.6.2 Δυνατότητες των Μεγάλων Γλωσσικών Μοντέλων

Η έκρηξη στην ανάπτυξη των Μεγάλων Γλωσσικών Μοντέλων (LLMs) συνοδεύτηκε από την επανάσταση που έφεραν στην NLP και στην τεχνητή νοημοσύνη. Τα LLMs έχουν πραγματοποιήσει αλματώδη πρόοδο, επιδεικνύοντας μια άνευ προηγουμένου ικανότητα στην κατανόηση, δημιουργία και εξαγωγή συμπερασμάτων από την ανθρώπινη γλώσσα επιδεικνύοντας αξιοσημείωτη ακρίβεια και ευελιξία [72]. Οι δυνατότητές τους εξελίσσονται συνεχώς, καθώς το μέγεθος των μοντέλων μεγαλώνει και νέες αναδυόμενες ικανότητές τους ανακλύπουν [54]. Παρακάτω, αναλύουμε κάποιες από τις βασικές τους δυνατότητες που τα καθιστούν μια δύναμη μεταμορφωτική ανακάλυψη.

### Εξαγωγή συμπερασμάτων (Inference)

Το inference (εξαγωγή συμπερασμάτων) είναι θεμελιώδης διαδικασία στα LLMs και η οποία, ουσιαστικά, επιτρέπει την αξιοποίηση των υπόλοιπων δυνατοτήτων τους, κάποιες εκ των οποίων



Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

περιγράφονται παρακάτω. Αντιπροσωπεύει την αντίθετη διαδικασία από τη διαδικασία εκπαίδευσης. Κατά τη διάρκεια της εκπαίδευσης, ένα μοντέλο μαθαίνει από ένα τεράστιο σύνολο δεδομένων, ώστε να καταγράψει την περιπλοκότητα της γλώσσας και των συμφραζόμενων. Αντίθετα, κατά τη διάρκεια του inference, ένας χρήστης εισάγει μια προτροπή (prompt) και το LLM εμπλέκεται σε μια διαδικασία παραγωγής αποκρίσεων. Αυτή η διαδικασία περιλαμβάνει την χρήση τα σταθερών προ-εκπαιδευμένων βαρών του μοντέλου, για να κατανοήσει το κείμενο εισόδου και να παράγει κείμενο ως έξοδο [86]. Η διαδικασία εξαγωγής συμπερασμάτων LLMs χωρίζεται σε δύο στάδια: το στάδιο προ-πλήρωσης (Prefill Stage) και το στάδιο αποκωδικοποίησης (Decode Stage)

*Prefill Stage (στάδιο προ-πλήρωσης):* Το αρχικό βήμα περιλαμβάνει την επεξεργασία της προτροπής για τη δημιουργία και αποθήκευση ζευγών κλειδιού-τιμής, γνωστά ως key-value cache (κρυφή μνήμη KV). Η κρυφή μνήμη βοηθά το μοντέλο να παράγει έξοδο, αποθηκεύοντας σχετικές πληροφορίες.

*Decode Stage (στάδιο αποκωδικοποίησης):* Εδώ συμβαίνει η πραγματική δημιουργία κειμένου. Το μοντέλο χρησιμοποιεί αποθηκευμένες πληροφορίες στην κρυφή μνήμη KV για τη δημιουργία εξόδου κειμένου. Το μοντέλο παράγει, βήμα – βήμα, tokens – λέξεις δηλαδή ή μέρη λέξεων. Η δημιουργία κάθε token επηρεάζεται από τα tokens που έχουν ήδη δημιουργηθεί, σχηματίζοντας με αυτόν τον τρόπο προτάσεις.

Έχει παρατηρηθεί όμως, ότι η δυνατότητα εξαγωγής συμπερασμάτων LLM περιορίζεται μερικές φορές από την εξάρτηση των μοντέλων σε μοτίβα στα δεδομένα και όχι στην αληθινή λογική συλλογιστική, βασίζεται δε στην ανάκληση των σχετικών απομνημονεύσεων από τα δεδομένα εκπαίδευσης, καθώς και σε προκαταλήψεις από τα δεδομένα εκπαίδευσης, όπως η συχνότητα εμφάνισης του όρου. Έτσι, μπορεί να οδηγηθούν σε σφάλματα κρίσης ή "παραισθήσεις" με κατασκευή πληροφοριών (hallucinations) [87].

Η ολοκλήρωση του inference εξαρτάται σε μεγάλο βαθμό από την υπολογιστική ισχύ και το μέγεθος της μνήμης του hardware που χρησιμοποιείται. Η παρακολούθηση των επιδόσεων του μοντέλου στο inference είναι εξαιρετικά χρήσιμη, καθώς επιτρέπει την αξιολόγηση της αποτελεσματικότητας και αξιοπιστίας LLM – κρίσιμα στοιχεία για τον προσδιορισμό της ικανότητάς του να αποδίδει σε πραγματικά σενάρια χρήσης και εντός αποδεκτού χρονικού πλαισίου. Η σωστή αξιολόγηση των LLMs αναφορικά με το inference βοηθά στην επιλογή του κατάλληλου μοντέλου για την περίπτωση χρήσης για την οποία προορίζεται, με βέλτιστη αξιοποίηση των διαθέσιμων πόρων. Δύο είναι τα βασικά μεγέθη για την μέτρηση της απόδοσης του inference: το latency (καθυστέρηση) και το throughput (ρυθμός επεξεργασίας). Latency είναι ο χρόνος που απαιτείται από το μοντέλο για την ολοκλήρωση της απόκρισης στο αίτημα, ενώ throughput είναι πόσα αιτήματα μπορεί να επεξεργαστεί ανά δεδομένη μονάδα χρόνου, π.χ. tokens/sec. Καθώς το inference είναι εξαιρετικά απαιτητικό σε ενέργεια, υπολογιστική ισχύ και μνήμη και για αυτό διάφορες τεχνικές αναπτύσσονται για βελτιστοποίηση των αναγκών αυτών, όπως το pruning.

### **Κατανόηση κειμένου (text understanding)**

Η κατανόηση κειμένου, στο πλαίσιο των LLMs, περιλαμβάνει την ικανότητα αυτών των μοντέλων να επεξεργάζονται, να ερμηνεύουν και να αντλούν νόημα από κείμενο φυσικής

γλώσσας. Η κατανόηση υπερβαίνει την απλή αναγνώριση λέξεων και φράσεων για να συμπεριλάβει την αντίληψη του πλαισίου (συμφραζόμενα), τη σημασιολογία (semantics), τον συναισθηματικό τόνο ακόμα και τις πολιτισμικές αποχρώσεις. Για τα LLMs η κατανόηση κειμένου είναι θεμελιώδης ικανότητα, επιτρέποντάς τους να δημιουργούν συνεκτικές απαντήσεις, να απαντούν σε ερωτήσεις με ακρίβεια και να εκτελούν ένα ευρύ φάσμα εργασιών (tasks) NLP που απαιτούν βαθιά κατανόηση της γλώσσας, όπως χρησιμοποιείται σε πραγματικές συνθήκες. Βασίζεται δε, στην εκπαίδευση των μοντέλων σε τεράστια σύνολα δεδομένων, όπως και η παραγωγή κειμένου [15].

Παρά τις σημαντικές προόδους, η κατανόηση κειμένου στα LLM αντιμετωπίζει αρκετές προκλήσεις λόγω της πολυπλοκότητας της γλώσσας. Για παράδειγμα, η γλώσσα είναι εγγενώς διφορούμενη και σε αρκετές περιπτώσεις οι λέξεις έχουν διαφορετική σημασία ανάλογα με τα συμφραζόμενα. Η ανίχνευση επίσης του τόνου του κειμένου (π.χ. ειρωνεία ή σαρκασμός) απαιτεί και την επίγνωση των κοινωνικών κανόνων και προσδοκιών, πέρα από την κυριολεκτική κατανόηση του κειμένου, μια πολύπλοκη διαδικασία για τα μοντέλα [75].

### ***Παραγωγή κειμένου (text generation)***

Η δημιουργία κειμένου, αναφερόμενη στα Μεγάλα Γλωσσικά Μοντέλα (LLM) παραπέμπει στη διαδικασία αυτόματης παραγωγής κειμένου, το οποίο προσομοιάζει τον ανθρώπινο λόγο, με βάση προτροπές εισαγωγής (prompts) ή δεδομένα. Τα LLMs αξιοποιούν την εκτεταμένη εκπαίδευσή τους σε τεράστια σύνολα δεδομένων για να δημιουργήσουν συνεκτικό, σχετικό με τα συμφραζόμενα κείμενο που εκτείνεται σε ένα ευρύ φάσμα στυλ και μορφών, που κυμαίνεται από απλές προτάσεις έως σύνθετες αφηγήσεις ή τεχνικές αναφορές. Βάση για την ικανότητα παραγωγής κειμένου είναι ο μηχανισμός αυτό-προσοχής (self-attention) που αναφέρθηκε παραπάνω. Αυτή τους η ικανότητα τα καθιστά ένα πολύ σημαντικό εργαλείο γενικής χρήσης, ακόμα και για εργασίες για τις οποίες δεν έχουν εκπαιδευτεί ειδικά [14, 55]. Οι εργασίες που μπορούν να εκτελούν καλύπτουν μεγάλη κλίμακα επιλογών από απλή γλωσσική μετάφραση και απαντήσεις σε ερωτήσεις, σε πιο σύνθετες εργασίες, όπως η σύνοψη (συμπύκνωση μεγάλων κειμένων σε συνοπτικές περιλήψεις, εργασία που περιλαμβάνει την κατανόηση των κύριων σημείων του κειμένου και τη διάκρισή τους από λιγότερο σημαντικές λεπτομέρειες), η δημιουργία κειμένου ακόμα και η δημιουργία κώδικα σε διάφορες γλώσσες προγραμματισμού [56]. Η χρησιμότητα των LLMs ενισχύεται περαιτέρω από την ικανότητά τους να προσαρμόζονται στο συγκεκριμένο στυλ και τόνο του κειμένου που επεξεργάζονται, καθιστώντας τα αποτελέσματα πιο φιλικά προς τον χρήστη και με επίγνωση των συμφραζόμενων [57].

### ***Text classification (Κατηγοριοποίηση κειμένου)***

Τα LLMs χρησιμοποιούν την ικανότητά τους στη γλωσσική κατανόηση για να κατηγοριοποιήσουν το κείμενο σε προκαθορισμένες τάξεις ή ετικέτες με βάση το περιεχόμενό του. Μοντέλα, όπως το BERT και όσα βασίζονται στην αρχιτεκτονική Transformer, μπορούν να κατανοούν πολύπλοκα γλωσσικά μοτίβα και αποχρώσεις. Ειδικά με το κατάλληλο finetuning, μπορούν να αποδώσουν αποτελεσματικά σε εργασίες κατηγοριοποίησης κειμένου.[88] Αυτή η δυνατότητα χρησιμοποιείται ευρέως στον εντοπισμό θεμάτων, το φιλτράρισμα ανεπιθύμητων μηνυμάτων και άλλες εφαρμογές.

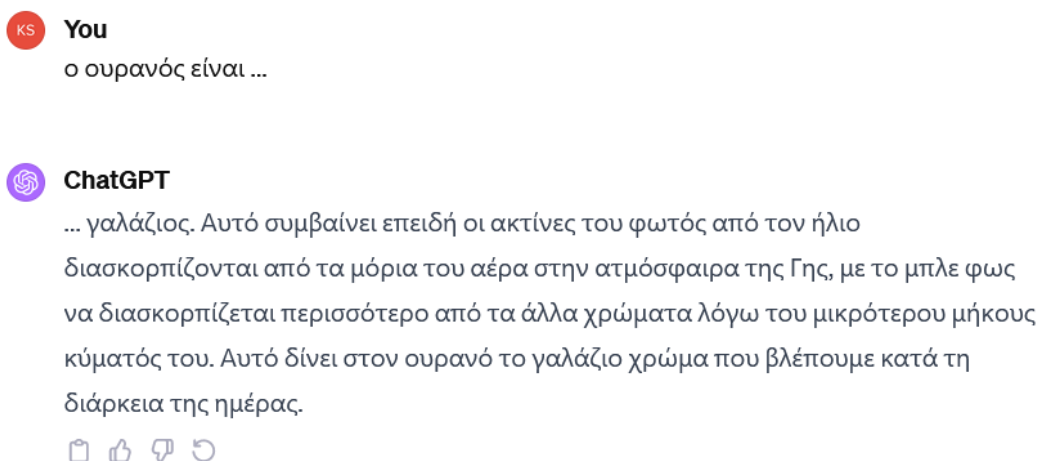
### **Ανάλυση συναισθήματος (sentiment analysis)**

Η ανάλυση συναισθήματος περιλαμβάνει τον προσδιορισμό του συναισθηματικού τόνου πίσω από μια σειρά λέξεων ώστε να κατανοήσουμε τις στάσεις, τις απόψεις και τα συναισθήματα που εκφράζονται π.χ. σε μια δημοσίευση σε μέσα κοινωνικής δικτύωσης, σε ένα έγγραφο ή μια πρόταση. Τα LLMs έχουν δείξει μια αξιοσημείωτη ικανότητα στο να ταξινομούν το κείμενο που ανάλογα με το εάν εκφράζει θετικά, αρνητικά ή ουδέτερα συναισθήματα, καθώς και να ανιχνεύουν πιο διαφοροποιημένες συναισθηματικές καταστάσεις ή εντάσεις.

## **2.7 Prompt engineering**

Με την εξέλιξη των Μεγάλων Γλωσσικών Μοντέλων (LLMs) , μοντέλα δηλαδή που είναι ικανά να παράγουν απαντήσεις με συνοχή και συνάφεια, αναδείχθηκε και η ανάγκη για την σχεδίαση και ανάλυση του βέλτιστου τρόπου επικοινωνίας με τα παραπάνω μοντέλα. Η διαδικασία αυτή ονομάζεται Prompt Engineering.

Το prompt (προτροπή) είναι το κείμενο που δίνεται στο μοντέλο, ώστε να ξεκινήσει να παράγει έξοδο. Με την προτροπή το μοντέλο καθοδηγείται προς το να εξάγει το επιθυμητό αποτέλεσμα και η τεχνική αυτή είναι καθοριστική για την βελτιστοποίηση των αποτελεσμάτων των LLMs. Ενώ ξεκίνησε σαν μια στοιχειώδη πρακτική για την καθοδήγηση του μοντέλου, έχει εξελιχθεί σε ένα πεδίο έρευνας με μεθοδολογίες και καλές πρακτικές. Ο στόχος βέβαια, παραμένει η παραγωγή μιας εξόδου με ακρίβεια, συνοχή και συνάφεια προς το ζητούμενο, καθώς και η εξερεύνηση του πλήρους βάθους των δυνατοτήτων των LLMs σε όλο εύρος των εφαρμογών τους. [16]. Παρακάτω παρατίθεται μια απλή μορφή προτροπής:



Σχήμα 16:Βασική Προτροπή

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

### 2.7.1 Ρυθμίσεις των LLMs

Στην περίπτωση που η αλληλεπίδραση με το LLM γίνεται με τη χρήση API, είναι δυνατόν να καθορίσει ο χρήστης διάφορες παραμέτρους, οι οποίες και παίζουν σημαντικό ρόλο στην επιθυμητή έξοδο του μοντέλου.

Παρακάτω αναφέρονται συνηθισμένες παράμετροι:

**Θερμοκρασία (Temperature):** Η ρύθμιση αυτή μπορεί να λάβει τιμές μεταξύ 0.2 και 2.0 και δηλώνει την τυχαιότητα της απάντησης. Δηλαδή, όσο μικρότερη η θερμοκρασία, το μοντέλο επιλέγει ως έξοδο το πιο πιθανό token, σε αντίθετη περίπτωση ενθαρρύνεται μια πιο τυχαία (και δημιουργική) απάντηση.

Οι πιθανές αποκρίσεις του μοντέλου περνούν από μια εκδοχή της Softmax συνάρτησης (που αναφέρθηκε νωρίτερα) και η οποία κανονικοποιεί τις πιθανότητες των αποκρίσεων, έτσι ώστε να έχουν άθροισμα 1. Μια εκδοχή της συνάρτησης Softmax με την θερμοκρασία  $T$  στον τύπο είναι η παρακάτω [89]:

$$\text{softmax}_T(x_i) = \frac{e^{\frac{x_i}{T}}}{\sum_{j=1}^n e^{\frac{x_j}{T}}}$$

Μεγαλώνοντας την τιμή της θερμοκρασίας αυξάνονται οι πιθανότητες των λιγότερο πιθανών token, διευρύνοντας έτσι το εύρος των πιθανών υποψηφίων για την πρόβλεψη του επόμενου token. Ουσιαστικά, η τιμή της θερμοκρασίας επηρεάζει τα βάρη των πιθανών tokens, άρα για εργασίες που απαιτείται ακρίβεια συνίσταται χαμηλή τιμή θερμοκρασίας, ενώ στην περίπτωση που είναι επιθυμητή μια πιο δημιουργική έξοδος, μεγαλύτερες τιμές.

**Top P:** Είναι μια τεχνική δειγματοληψίας που ονομάζεται nucleus sampling (πυρηνική δειγματοληψία) και με την οποία ελέγχεται πόσο ντετερμινιστικό είναι το μοντέλο. Εάν επιλεγεί μικρή τιμή της παραμέτρου, το μοντέλο επιλέγει την απόκρισή του από μια μικρότερη δεξαμενή πιθανών επιλογών λέξεων (αυτών με την μεγαλύτερη πιθανότητα). Η γενική σύσταση είναι να αλλάζει η θερμοκρασία ή το Top P, αλλά όχι και τα δύο.

**Max Length:** Η παράμετρος αυτή αναφέρεται στον μέγιστο αριθμό από tokens τον οποίο μπορεί να εξάγει το μοντέλο στην απάντησή του. Με την ρύθμιση αυτή είναι δυνατόν να αποφευχθούν μακροσκελείς αποκρίσεις, αλλά και να ελεγχθεί το πιθανό κόστος χρήσης.

**Stop Sequences:** Το stop sequence είναι μια συμβολοσειρά για να σταματήσει το μοντέλο την παραγωγή tokens. Είναι ένας άλλος τρόπος, δηλαδή, για τον έλεγχο του μήκους και της δομής της απόκρισης του μοντέλου. Για παράδειγμα, αν έχει ζητηθεί η δημιουργία λιστών από το μοντέλο, καθορίζοντας το «11» ως stop sequence δηλώνεται στο μοντέλο να μην δημιουργεί λίστες με πάνω από 10 αντικείμενα. [20]

### 2.7.2 Βασικά στοιχεία του prompt engineering

Μια προτροπή περιλαμβάνει οποιοδήποτε από τα παρακάτω στοιχεία (όχι απαραίτητα όλα):

**Οδηγία :** μια συγκεκριμένη εργασία ή οδηγία που θέλουμε να εκτελέσει το μοντέλο

**Πλαίσιο συζήτησης :** επιπλέον πληροφορίες για την κατεύθυνση του μοντέλου σε καλύτερες αποκρίσεις

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

**Δεδομένα εισόδου:** την ερώτηση ή το θέμα για το οποίο θέλουμε απόκριση

**Οδηγός εξόδου:** το είδος ή η μορφή της εξόδου.

Για παράδειγμα, μια απλή προτροπή για κατηγοριοποίηση κειμένου (text classification) είναι :

Κατηγοριοποίησε το παρακάτω κείμενο σε ουδέτερο, αρνητικό ή θετικό.

Κείμενο: " η ταινία ήταν καλή"

Συναίσθημα:

**Απάντηση:**

Το κείμενο "η ταινία ήταν καλή" εκφράζει ένα θετικό συναίσθημα.

Στο παραπάνω παράδειγμα, η οδηγία είναι «Κατηγοριοποίησε το παρακάτω κείμενο σε ουδέτερο, αρνητικό ή θετικό», τα δεδομένα εισόδου με τις επιπλέον πληροφορίες το : «Κείμενο: " η ταινία ήταν καλή"» και τέλος, ο οδηγός εξόδου το «Συναίσθημα: ». Πλαίσιο συζήτησης δεν δόθηκε σε αυτό το παράδειγμα.

### 2.7.3 Γενικές οδηγίες/στρατηγικές για την δημιουργία προτροπών

Στο πλαίσιο της εξάπλωσης της χρήσης των LLMs ως διαλογικών συστημάτων, έχουν διαμορφωθεί οδηγίες για την δημιουργία των προτροπών, με σκοπό την πιο αποτελεσματική χρήση των LLMs και την καθοδήγηση των μοντέλων στην επιθυμητή έξοδο [18].

*Σαφείς οδηγίες*

Τα μοντέλα ανταποκρίνονται καλύτερα όταν λαμβάνουν ξεκάθαρες και σαφείς οδηγίες και δεν χρειάζεται να μαντέψουν τι ζητά ο χρήστης. Καθώς τα μοντέλα έχουν εκπαιδευθεί σε πληθώρα δεδομένων, μια ασαφής και γενική οδηγία μπορεί να οδηγήσει σε έξοδο η οποία να έχει νόημα σε κάποιο πλαίσιο, αλλά να μην είναι η βέλτιστη για άλλες περιπτώσεις. Στην περίπτωση που δίνονται συγκεκριμένες και σαφείς κατευθύνσεις, μειώνεται η αβεβαιότητα για το μοντέλο και οδηγείται στην επιθυμητή απόκριση. [16]

Στο πλαίσιο αυτό καλές πρακτικές είναι :

*Να δίνεται κείμενο αναφοράς. Π.χ.*

Εάν δεν γνωρίζεις την απάντηση, γράψε 'Δεν γνωρίζω την απάντηση'

*Να δίνονται σαφείς οδηγίες για την δομή της εξόδου. Π.χ.*

Γράψε λίστα από χώρες της Ευρωπαϊκής ένωσης. Η λίστα να είναι αριθμημένη.

*Να αποφεύγεται η ανακρίβεια και η αοριστία. Π.χ.*

Γράψε μια περίληψη 3-4 γραμμών από το παρακάτω κείμενο.

Η παραπάνω προτροπή είναι προτιμότερη από προτροπή της μορφής:

Γράψε μια περίληψη του παρακάτω κειμένου.

*Τοποθέτηση των οδηγιών στην αρχή της προτροπής και χρήση των ### ή "" για τον διαχωρισμό των οδηγιών και του κειμένου. Π.χ.*

Συνόψισε το παρακάτω κείμενο σε λίστα με τα πιο σημαντικά σημεία

Κείμενο : ""

{κείμενο}

""

*Διάσπαση πολύπλοκων εργασιών σε μικρότερες .[17] Π.χ.*

## Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Πρώτον, δώσε μια επισκόπηση. του θέματος. Δεύτερον, ανάφερε σε αριθμημένη λίστα τους κύριους ερευνητές στο πεδίο αυτό. Τρίτο, ανάφερε επιγραμματικά τις τελευταίες εξελίξεις.

### 2.7.4 Τεχνικές δημιουργίας προτροπών (prompts)

Ένας από τους μεγαλύτερους προβληματισμούς αναφορικά με το Prompt engineering είναι η δυσκολία σχεδίασης προτροπών για συγκεκριμένα είδη εργασιών και η έλλειψη αυτοματοποιημένων τρόπων για να γίνει η σχεδίαση. Γενικευμένες προτροπές είναι συνήθως λιγότερο αποτελεσματικές από άλλες που έχουν σχεδιαστεί για συγκεκριμένη εργασία. Στο πλαίσιο της έρευνας για την δημιουργία αυτοματοποιημένων σχετικών μεθόδων τρεις τεχνικές είναι, για την ώρα, αυτές που χρησιμοποιούνται:

#### 2.7.4.1.1 Zero-shot prompting

Η πρώτη τεχνική που αναφέρεται είναι η zero-shot prompting, η οποία βασίζεται στη δυνατότητα των Μεγάλων Γλωσσικών Μοντέλων να ανταποκρίνονται σε εργασίες ή ερωτήσεις για τις οποίες δεν έχουν συγκεκριμένα εκπαιδευτεί, χάρη στην εκτενή εκπαίδευσή τους σε μεγάλο όγκο δεδομένων. Αντί να παρέχονται συγκεκριμένα παραδείγματα, το μοντέλο επικαλείται την προϋπάρχουσα γνώση του για να επιλύσει την εκάστοτε εργασία. Έχει διαπιστωθεί ότι μοντέλα τα οποία έχουν υποστεί fine-tuning με οδηγίες, γνωστό ως “instruction tuning”, όπως το Mistral 7b instruct, επιδεικνύουν επαρκή απόκριση σε τέτοιες προκλήσεις [19]. Επιπλέον, σχετικές έρευνες [21] προτείνουν ότι με την (επόμενη) τεχνική του few-shot prompting το μοντέλο δεν μαθαίνει νέες εργασίες, αλλά εντοπίζει εργασίες που έχει ήδη μάθει.

#### 2.7.4.1.2 Few-shot prompting

Η επόμενη τεχνική είναι η few-shot prompting η οποία επιτρέπει την βελτίωση των αποκρίσεων των LLMs με την προσθήκη στο κείμενο της προτροπής μερικών υποδειγμάτων απόκρισης [15]. Αυτή η μέθοδος αξιοποιεί την μεγάλη γνωσιακή βάση των Μοντέλων και τις ικανότητες που έχουν αποκτήσει από την εκπαίδευσή τους τα μοντέλα, «θυμίζοντάς» τους μια εργασία που έχουν ήδη μάθει. [21]. Παρακάτω βλέπουμε μια προτροπή δομημένη με την μέθοδο few-shot

Κείμενο: "Είμαι πολύ χαρούμενος σήμερα!"

Συναίσθημα: Χαρά

Κείμενο: "Αισθάνομαι πολύ άσχημα"

Απάντηση :

Συναίσθημα: Λύπη ή Δυσαρέσκεια

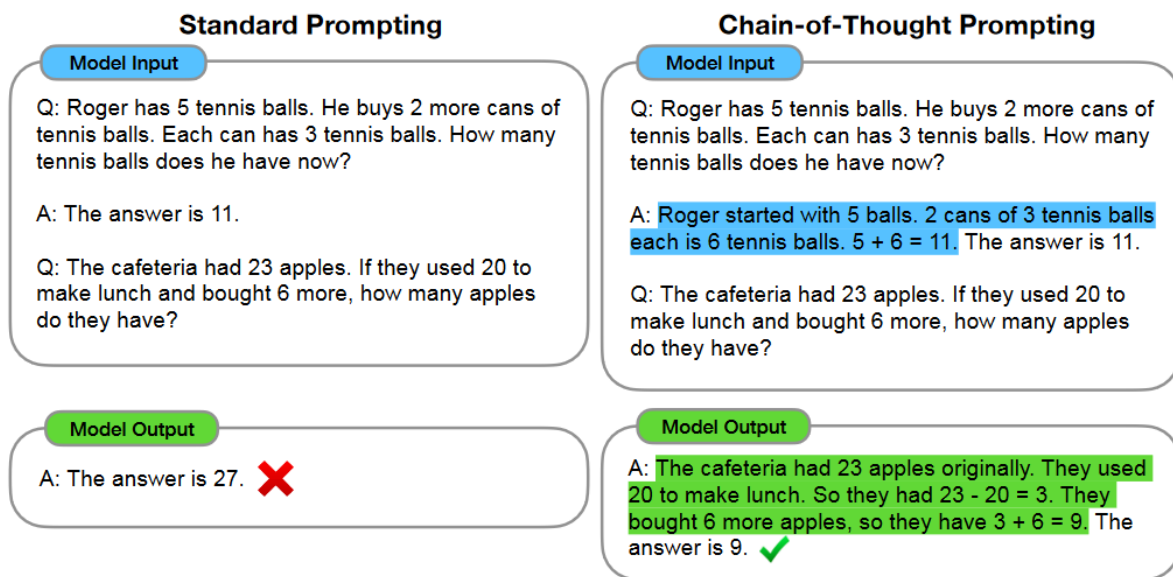
Παρόλο το επιτυχημένο παράδειγμα που αναφέρθηκε, η τεχνική few-shot έχει περιορισμούς και δεν είναι πάντα η ενδεδειγμένη. Η αποτελεσματικότητά της εξαρτάται από διάφορους παράγοντες όπως π.χ. το είδος και η ποιότητα των παραδειγμάτων που δίνονται καθώς και το είδος και η πολυπλοκότητα της εργασίας που ζητείται από το μοντέλο. Τα μοντέλα δεν είναι ικανά πάντα να γενικεύουν σωστά σε νέες, εργασίες ή δεδομένα που διαφέρουν σημαντικά από

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

αυτά στα οποία έχουν εκπαιδευθεί. Στις περιπτώσεις αυτές θα μπορούσε να χρησιμοποιηθεί η επόμενη τεχνική ή και να εξεταστεί η δυνατότητα finetuning του μοντέλου.

#### 2.7.4.2 Chain of Thought

Μια ακόμα τεχνική που έχει προταθεί για την καθοδήγηση των μοντέλων στην επιθυμητή έξοδο είναι η Chain-of-Thought. Στην προσέγγιση αυτή δεν ζητείται απευθείας η τελική απάντηση, αλλά αυτή προκύπτει μετά από ενδιάμεσα βήματα που οδηγούν στην λύση [51]. Συνδυάζεται συνήθως με την few-shot, δίνοντας στο μοντέλο κάποιο-α παράδειγμα-τα από το ζητούμενο, όπως φαίνεται στην παρακάτω εικόνα:



Σχήμα 17 – Παράδειγμα προτροπής με βάση την τεχνική Chain-of-Thought [σχήμα από 51]

Η chain-of-thought είναι ιδιαίτερα χρήσιμη π.χ. σε μαθηματικά ή λογικά προβλήματα, που απαιτούν αρκετά βήματα για να επιλυθούν ή σε περιπτώσεις που η λογική ακολουθία της επίλυσης είναι σημαντική όπως και η λύση.

## 2.8 Περιοχές εφαρμογής

Τα LLMs έχουν χρησιμοποιηθεί σε διάφορους τομείς, επιδεικνύοντας τις ικανότητές και την προσαρμοστικότητά τους, διαφοροποιώντας τον τρόπο που πραγματοποιούνται μια πληθώρα εργασιών.

Για παράδειγμα, τα LLMs αρχίζουν να δείχνουν τις δυνατότητές τους στον μετασχηματισμό του **ιατρικού χώρου**. Η ικανότητα των μοντέλων να παράγουν απαντήσεις κειμένου με συνάφεια και ακρίβεια, όπως π.χ. των αποτελεσμάτων τους στις εξετάσεις για απόκτηση άδειας ασκήσεως ιατρικής των Ηνωμένων Πολιτειών (USMLE) με επίπεδο αντίστοιχο φοιτητή ιατρικής 3<sup>ου</sup> έτους [58], υποδηλώνει έναν πολλά υποσχόμενο δρόμο για την ενίσχυση τόσο των κλινικών όσο και των εκπαιδευτικών πτυχών στην ιατρική.

Σε κλινικά περιβάλλοντα, LLMs έχουν χρησιμοποιηθεί για την εξαγωγή σημαντικών πληροφοριών για τους ασθενείς από κλινικές σημειώσεις, βελτιώνοντας την αποτελεσματικότητα και την ακρίβεια της τήρησης ιατρικών αρχείων [44]. Έχουν δείξει ήδη αξιοσημείωτες ικανότητες, ιδιαίτερα στη διάγνωση και τη φροντίδα ασθενών. Αναλύοντας ερωτήματα ασθενών και κλινικές σημειώσεις, αυτά τα μοντέλα μπορούν να παρέχουν προκαταρκτικές διαγνώσεις, να προτείνουν πιθανές οδούς θεραπείας και ακόμη και να παράγουν ιατρικές συμβουλές για τον ασθενή. Ωστόσο, είναι σημαντικό να σημειωθεί ότι, ενώ αυτά τα μοντέλα παρουσιάζουν ικανοποιητικά επίπεδα απόδοσης, δεν επιδεικνύουν, στην παρούσα φάση, τέτοιο βαθμό ακρίβειας και αποτελεσματικότητας ώστε να μπορούν να χρησιμοποιηθούν αυτόνομα και χωρίς ιατρική επίβλεψη [60, 71].

Μια άλλη σημαντική εφαρμογή είναι στην ιατρική εκπαίδευση, όπου τα LLMs μπορούν να χρησιμεύσουν ως εξελιγμένα συστήματα διδασκαλίας. Έχουν τη δυνατότητα να βοηθήσουν τους φοιτητές ιατρικής στην προετοιμασία τους για τις εξετάσεις παρέχοντας υλικό μελέτης, απαντώντας σε σύνθετες ιατρικές ερωτήσεις και προσομοιώνοντας περιπτώσεις ασθενών για κλινική πρακτική. Αυτό θα μπορούσε να φέρει επανάσταση στον τρόπο με τον οποίο παρέχεται η ιατρική εκπαίδευση, προσφέροντας μια πιο εξατομικευμένη και διαδραστική εμπειρία μάθησης. [59].

Αντίστοιχου τύπου εφαρμογή έχουν τα LLMs και στο **νομικό** τομέα. Για παράδειγμα, μπορούν να προσφέρουν σημαντική βοήθεια στο κομμάτι της νομικής έρευνας, αντλώντας με ταχύτητα σχετική νομολογία, καταστατικά, κανονισμούς και νομικά προηγούμενα από τεράστιες νομικές βάσεις δεδομένων. Αυτό επιταχύνει τη διαδικασία της έρευνας και βοηθά τους δικηγόρους να προετοιμάσουν τις υποθέσεις τους πιο αποτελεσματικά. Επιπλέον, με την εκπαίδευσή τους σε ειδικά εξειδικευμένα δεδομένα σχετικά με την νομοθεσία, είναι ικανά να παράγουν νομική επιχειρηματολογία [73] και να απαντήσουν σε νομικές ερωτήσεις [74], αποτελώντας ένα πολύτιμο εργαλείο.

Ομοίως, στον τομέα των **χρηματοοικονομικών**, η χρήση των LLMs έχει να επιδείξει σημαντικά οφέλη με διάφορες εφαρμογές. Για παράδειγμα, μπορούν να συνδράμουν στην ανίχνευση ύποπτων δραστηριοτήτων, αφού έχουν την δυνατότητα να αναλύουν μεγάλους όγκους οικονομικών δεδομένων σε πραγματικό χρόνο Έτσι, εντοπίζοντας ασυνήθιστα μοτίβα ή ανωμαλίες, ενισχύουν την ασφάλεια των συναλλαγών και προστατεύουν από οικονομικές απάτες.

Η ικανότητά τους να αναλύουν και να ερμηνεύουν τεράστιο όγκο δεδομένων κειμένου, όπως οικονομικές εκθέσεις, οικονομικούς δείκτες και εταιρικά αρχεία, έχει προκαλέσει δομικές αλλαγές στον χρηματοπιστωτικό τομέα. Τα LLMs μπορούν να αφομοιώσουν πολύπλοκα χρηματοοικονομικά έγγραφα, εντοπίζοντας βασικές τάσεις και ιδέες που μπορεί να μην είναι άμεσα εμφανείς στους ανθρώπινους αναλυτές. Με τον τρόπο αυτό μπορούν να συνδράμουν αποφασιστικά σε διάφορες εργασίες χρηματοοικονομικής πρόβλεψης και ανάλυσης, συμπεριλαμβανομένης της πρόβλεψης των κινήσεων της αγοράς, της αξιολόγησης του πιστωτικού κινδύνου και της αξιολόγησης της απόδοσης του χαρτοφυλακίου [38].

Τα LLMs χρησιμοποιούνται στην ανάλυση νέων και αναφορών σε μέσα κοινωνικής δικτύωσης, ακόμα και την ανάλυση του συναισθήματος στις χρηματοοικονομικές ειδήσεις, παρέχοντας πολύτιμες γνώσεις για την πρόβλεψη των τάσεων της αγοράς και τη διαμόρφωση επενδυτικών στρατηγικών σε πραγματικό χρόνο. [45]. Με την επεξεργασία αυτού του



διαφορετικού συνόλου δεδομένων, μπορούν να εντοπίσουν μοτίβα και σημάδια που προηγούνται των τάσεων της αγοράς, προσφέροντας στους επενδυτές πολύτιμες πληροφορίες σχετικά με πιθανές ανοδικές ή πτωτικές συνθήκες της αγοράς. Οι πληροφορίες μπορούν να βοηθήσουν τους διαχειριστές χαρτοφυλακίων να λαμβάνουν πιο ενημερωμένες αποφάσεις, οδηγώντας ενδεχομένως σε υψηλότερες αποδόσεις των επενδύσεων [61]. Αντίστοιχα, τα LLMs μπορούν να προσομοιάσουν διαφορετικά οικονομικά σενάρια και τον πιθανό αντίκτυπό τους στην απόδοση του χαρτοφυλακίου, επιτρέποντας στους επενδυτές να προετοιμαστούν και να μετριάσουν τους κινδύνους που σχετίζονται με την αστάθεια της αγοράς. Αυτή η εφαρμογή των μεγάλων γλωσσικών μοντέλων, στην ανάλυση σεναρίων και στις προσομοιώσεις ακραίων καταστάσεων, είναι εξαιρετικά σημαντική για τον στρατηγικό σχεδιασμό και τη διαχείριση κινδύνου σε χρηματοοικονομικά χαρτοφυλάκια [62].

Ωστόσο, και παρά τις επιτυχημένες προσπάθειες, η προσαρμογή των LLMs σε ειδικές απαιτήσεις ανά τομέα παρουσιάζει αξιοσημείωτες προκλήσεις, όπως για παράδειγμα την μείωση των προκαταλήψεων (biases) που βρίσκονται ενσωματωμένες στα σύνολα δεδομένων εκπαίδευσης και οι οποίες μπορούν να παραμορφώσουν τα αποτελέσματα των μοντέλων και τις διαδικασίες λήψης αποφάσεων [46, 66].

## 2.9 Ανάλυση Αγοράς Εργασίας

Όσον αφορά τον τομέα ενδιαφέροντος αυτής της μελέτης, την αγορά εργασίας, η ανάπτυξη των LLMs έφερε νέες δυνατότητες στη ανάλυση των δεδομένων της. Οι πληροφορίες που ενδιαφέρουν τους αναλυτές, όπως π.χ. οι αγγελίες εργασίας ώστε να αναγνωριστούν οι τάσεις στα επαγγέλματα τα οποία έχει ανάγκη η αγορά εργασίας, είναι δεδομένα τεράστια σε μέγεθος και συνήθως σε αδόμητη μορφή. Για αυτό, η στροφή προς προηγμένες τεχνικές NLP για την εξαγωγή πληροφοριών από αγγελίες εργασίας και βιογραφικά είναι μια σημαντική εξέλιξη. Η έρευνα έχει επισημάνει τις βαθιές δυνατότητες των LLMs να εμπλουτίσουν την κατανόησή μας για τις τάσεις της αγοράς εργασίας και τις απαιτήσεις δεξιοτήτων [24, 39]. Ο τομέας, ωστόσο, βρίσκεται ακόμα στα αρχικά του στάδια, προσφέροντας ευκαιρίες για μελλοντικές μελέτες για την αξιολόγηση της αποτελεσματικότητας των διαφόρων LLMs στη διάκριση λεπτομερών πληροφοριών σε ετερογενείς αγορές εργασίας [50].

Μια σημαντική πηγή πληροφοριών για την αγορά εργασίας είναι η μελέτη των αγγελιών εργασίας. Μια σειρά από μεθοδολογίες εξαγωγής πληροφοριών, από την απλή εξαγωγή λέξεων-κλειδιών έως την εξελιγμένη βαθιά σημασιολογική ανάλυση, έχουν χρησιμοποιηθεί για την εξόρυξη βασικών δεδομένων από αυτές, όπως οι απαιτούμενες δεξιότητες, τα προσόντα και οι περιγραφές θέσεων εργασίας [36]. Με βάση αυτές τις μελέτες έχουμε ως στόχο για την μετέπειτα ανάλυσή μας να αξιολογήσουμε πώς τα LLMs μπορούν να βελτιώσουν τη λεπτομέρεια και την ακρίβεια των πληροφοριών που προέρχονται από τις αγγελίες εργασίας, προσφέροντας με αυτόν τον τρόπο μια πλουσιότερη κατανόηση της δυναμικής της αγοράς εργασίας [37].

# 3

## *Μεθοδολογία*

Σκοπός της εργασίας είναι η αξιοποίηση των τεχνολογικών εξελίξεων στον τομέα των διαλογικών συστημάτων τεχνητής νοημοσύνης για την εξαγωγή πληροφοριών από αγγελίες εργασίας.

Πιο αναλυτικά, στο πλαίσιο της εργασίας πραγματοποιήθηκε μια πρακτική αξιολόγηση για να εκτιμηθεί πόσο αποτελεσματικά τα τελευταίας τεχνολογίας LLMs μπορούν να εξάγουν σημασιολογικά σημαντικές πληροφορίες από ένα μη επεξεργασμένο σύνολο από αγγελίες θέσεων εργασίας. Οι αγγελίες αυτές αποτελούνται από δεδομένα του έτους 2020, συγκεντρώθηκαν από δημοφιλείς πλατφόρμες αγγελιών εργασίας και διατηρήθηκαν σκοπίμως στην αρχική της μορφή. Αυτή η προσέγγιση στόχευε στην προσομοίωση του πραγματικού περιβάλλοντος δεδομένων που θα απαντούν τα LLM σε πραγματικά σενάρια χρήσης.

Στην μελέτη χρησιμοποιούνται μια σειρά από LLMs τα οποία και επιλέχθηκαν με βάση τις ικανότητές τους στη γλωσσική επεξεργασία, καθώς τις απαιτήσεις τους σε υπολογιστική ισχύ, ώστε να είναι δυνατή η χρήση τους σε υλικό (hardware) επιπέδου καταναλωτή (consumer-grade). Τα μοντέλα, gpt-3.5-turbo, llama-2-7b-chat.ggmlv3.q4\_0.bin, mistral-7b-instruct-v0.1.Q4\_0.gguf και orca-mini-3b-gguf2-q4\_0.gguf, είναι αυτά που επιλέχθηκαν μετά από μια συνολική διαδικασία συγκριτικής αξιολόγησης.

Αυτή η αξιολόγηση εξασφάλισε την καταλληλότητά τους για τυπικούς προσωπικούς υπολογιστές, την αξιολόγηση της γλωσσικής ακρίβειας, της λειτουργικής αποτελεσματικότητας και της ανταπόκρισης εντός τυπικών περιορισμών υπολογιστικών πόρων.

Στόχος είναι να εξαχθούν τα παρακάτω στοιχεία από τις περιγραφές αγγελιών του συνόλου δεδομένων :

- Job Title (τίτλος της θέσης εργασίας),

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

- Industry (η οικονομική δραστηριότητα που συνδέεται με την θέση εργασίας),
- Location (τόπος της εργασίας),
- Soft skills (οριζόντιες/ήπιες δεξιότητες - Μη τεχνικές δεξιότητες που περιλαμβάνουν τον τρόπο αλληλεπίδρασης με τους συναδέλφους, την επίλυση προβλημάτων και τη διαχείριση της εργασίας. Είναι χαρακτηριστικά γνωρίσματα και διαπροσωπικές δεξιότητες που χαρακτηρίζουν τις σχέσεις με άλλους ανθρώπους και συμπληρώνουν τις γνώσεις του εργαζόμενου στο χώρο εργασίας. Μερικά παραδείγματα περιλαμβάνουν επικοινωνία, ομαδική εργασία, προσαρμοστικότητα, ηγεσία, επίλυση προβλημάτων, ηθική εργασίας, διαχείριση χρόνου και συναισθηματική νοημοσύνη)
- Hard skills (κάθετες/σκληρές δεξιότητες - τεχνικές δεξιότητες και γνώσεις, σχετικές με το αντικείμενο εργασίας που αποκτώνται μέσω σπουδών, εργασιακής εμπειρίας κλπ. Για παράδειγμα γνώση ξένων γλωσσών, γλωσσών προγραμματισμού κ.α.)

### 3.1 Συλλογή Δεδομένων

Για την συλλογή δεδομένων χρησιμοποιήθηκαν τεχνικές web crawling (ανίχνευσης ιστού), ειδικότερα με την βοήθεια εργαλείων όπως το BeautifulSoup και το Scrapy, ώστε να συλλεχθούν αγγελίες εργασίας από διάφορες διαδικτυακές πλατφόρμες με αυτοματοποιημένο τρόπο, όπως περιγράφονται στο [22]. Το dataset (σύνολο δεδομένων) που προέκυψε υποβλήθηκε σε διεξοδικές διαδικασίες για να διασφαλιστεί η ακρίβεια και η αξιοπιστία του, συμπεριλαμβανομένης της διόρθωσης σφαλμάτων, της αφαίρεσης διπλότυπων, της εξάλειψης ειδικών χαρακτήρων και της εξαίρεσης άσχετου περιεχομένου. Χρησιμοποιήθηκαν μη αυτόματες μέθοδοι καθαρισμού και επιβεβαίωσης σε συνδυασμό με Python scripts, ώστε να επιλυθούν τυχόν ζητήματα στα δεδομένα. Το ολοκληρωμένο dataset περιλαμβάνει περισσότερες από 10.000 αγγελίες από 5 αξιόπιστες διαδικτυακές πηγές παρέχοντας μια πλήρη επισκόπηση της αγοράς εργασίας κατά τη διάρκεια αυτού του συγκεκριμένου χρονικού πλαισίου.

Στην παρούσα εργασία αξιοποιήθηκαν 3005 καταχωρίσεις από το παραπάνω σύνολο δεδομένων. Εξετάζοντας τα δεδομένα που χρησιμοποιήθηκαν, ξενοδοχεία και παρόμοια καταλύματα αποτελούν την πλειοψηφία των καταχωρίσεων, ακολουθούμενα από έναν συγκριτικά μικρότερο αριθμό καταχωρίσεων που σχετίζονται με εστιατόρια και υπηρεσίες τροφοδοσίας. Αυτές οι βιομηχανίες αναδεικνύονται ως τα κύρια σημεία εστίασης οικονομικού ενδιαφέροντος, υποδηλώνοντας την επέκταση του τουριστικού τομέα, παρά το ότι συλλέχθηκαν κατά την περίοδο της πανδημίας. Επιπλέον, το σύνολο δεδομένων περιλαμβάνει αγγελίες που σχετίζονται με δραστηριότητες ταξιδιωτικών πρακτορείων και θέσεις εργασίας σε κτηματομεσιτικά γραφεία και διαχείριση (management), αν και σε μικρότερο βαθμό.

Εξετάζοντας τους επαγγελματικούς κωδικούς στις αγγελίες, διαπιστώσαμε ότι περιέχονται συνολικά 159 διακριτοί κωδικοί. Σε συμφωνία με τις προηγούμενες αναλύσεις, η πλειονότητα των αγγελιών θέσεων εργασίας επικεντρώνεται στον κλάδο των υπηρεσιών, ο οποίος αντανακλά την οικονομική δομή της Ελλάδας, όπου ο τουρισμός και οι συναφείς υπηρεσίες διαδραματίζουν κρίσιμο ρόλο. Συγκεκριμένα, ένας σημαντικός αριθμός θέσεων εργασίας αφορά επαγγέλματα όπως σερβιτόροι, μάγειρες, υπεύθυνοι υποδοχής ξενοδοχείων, μάρμεν, καθαρίστριες, σεφ, βοηθοί πωλήσεων και βοηθοί κουζίνας. Η επικράτηση αυτών των θέσεων εργασίας δείχνει ότι οι τομείς αυτοί παρουσιάζουν ανθεκτικότητα ή και ανάκαμψη, πιθανώς λόγω της αύξησης του

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

εγγώριου τουρισμού ή της επέκτασης υπηρεσιών, όπως παράδοση στο σπίτι, στην εστίαση. Η παρουσία κτηματομεσιτών στη λίστα των πιο συχνών αγγελιών θέσεων εργασίας είναι επίσης σημαντική και θα μπορούσε να αποδοθεί στη συνεχιζόμενη δραστηριότητα στην αγορά ακινήτων, δυναμικά τροφοδοτούμενη και από αλλαγές στις προτιμήσεις στέγασης ή σε επενδύσεις κατά τη διάρκεια της πανδημίας.

## 3.2 Επιλεγμένα μοντέλα και εφαρμογές

Στην επιλογή των μοντέλων ρόλο έπαιξαν οι διαφορές στην απόδοση μεταξύ των εμπορικών LLMs και αυτών ανοιχτού κώδικα. Συγκεκριμένα, οι συγκριτικές αναλύσεις μεταξύ εμπορικών (κλειστού κώδικα) και ανοιχτού κώδικα LLMs έχουν περιγράψει σαφείς διαφορές όσον αφορά στην απόδοση και στην προσβασιμότητα [34]. Τα εμπορικά LLMs χαρακτηρίζονται συνήθως από τα εξελιγμένα χαρακτηριστικά τους και τις ολοκληρωμένες υπηρεσίες υποστήριξης που καλύπτουν ένα ευρύ φάσμα πολύπλοκων εφαρμογών [67]. Αντίθετα, τα LLM ανοιχτού κώδικα διακρίνονται για τη διαφάνεια και την προσαρμοστικότητά τους, επιτρέποντας εκτεταμένη προσαρμογή στις απαιτήσεις μας και διευκολύνοντας τη διαφάνεια της έρευνας [47,48]. Χρησιμοποιώντας ως βάση αυτή την αντίθεση για τη συγκριτική μελέτη που παρουσιάζουμε, επιδιώκουμε να δείξουμε πώς αποδίδουν τα διαφορετικά LLM στην ανάλυση της αγοράς εργασίας, παρέχοντας έτσι πληροφορίες για τη βέλτιστη επιλογή μοντέλων με βάση συγκεκριμένες ανάγκες ανάλυσης [35].

### 3.2.1 GPT-3.5-turbo

Το **GPT-3.5-turbo** μοντέλο αναπτύχθηκε από την εταιρεία OpenAI ως εξέλιξη των προηγούμενων μοντέλων του ίδιου οργανισμού (το GPT-2 και το GPT-3). Ανήκει στην ομάδα μοντέλων GPT-3.5 και είναι ειδικά προσαρμοσμένο για συνομιλία. Έχει όριο τα 4,096 tokens, ενώ δεν έχει ανακοινωθεί επίσημα ο αριθμός των παραμέτρων του. (το προηγούμενο μοντέλο GPT-3 έχει 175 δισεκατομμύρια παραμέτρους).[25, 26]

Τα μοντέλα συνομιλίας λαμβάνουν μια λίστα μηνυμάτων ως είσοδο και επιστρέφουν ένα μήνυμα που δημιουργείται από το μοντέλο ως έξοδο. Παρόλο που η μορφή συνομιλίας έχει σχεδιαστεί για να διευκολύνει τις συνομιλίες με έναν αριθμό ερωτήσεων - απαντήσεων, είναι εξίσου χρήσιμη για απλές εργασίες μίας απάντησης, χωρίς καμία συνομιλία.

Οι υπηρεσίες του μοντέλου είναι προσβάσιμες χρησιμοποιώντας το Chat Completion API [41].

Μια κλήση του Chat Completions API είναι όπως φαίνεται παρακάτω :

```
from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
    model="gpt-3.5-turbo",
    messages= [
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers won the World Series in 2020."},
        {"role": "user", "content": "Where was it played?"}
```

```
]
)
```

\*το παράδειγμα είναι από τις επίσημες οδηγίες της OpenAI

Η κύρια είσοδος είναι η μεταβλητή μηνυμάτων. Τα μηνύματα πρέπει να είναι μια σειρά (array) από αντικείμενα μηνυμάτων (message objects), όπου κάθε αντικείμενο έχει έναν ρόλο (είτε "σύστημα", "χρήστης" ή "βοηθός") και περιεχόμενο. Οι συνομιλίες μπορεί να περιέχουν αρκετές ερωτω-απαντήσεις ή να είναι σύντομες (με ένα μήνυμα).

Συνήθως, μια συνομιλία μορφοποιείται πρώτα με ένα μήνυμα συστήματος, ακολουθούμενο από εναλλασσόμενα μηνύματα χρήστη και βοηθού.

Το μήνυμα συστήματος βοηθά στη ρύθμιση της συμπεριφοράς του βοηθού, π.χ. να δοθούν συγκεκριμένες οδηγίες σχετικά με τον τρόπο συμπεριφοράς του σε όλη τη διάρκεια της συνομιλίας. Ωστόσο, το μήνυμα συστήματος είναι προαιρετικό και η συμπεριφορά του μοντέλου χωρίς μήνυμα συστήματος είναι πιθανό να είναι παρόμοια με τη χρήση ενός γενικού μηνύματος, όπως "Είστε χρήσιμος βοηθός".

Τα μηνύματα χρήστη παρέχουν αιτήματα ή σχόλια για να απαντήσει ο βοηθός. Τα μηνύματα του Βοηθού αποθηκεύουν προηγούμενες απαντήσεις του βοηθού, αλλά μπορούν επίσης να γραφτούν από τον χρήστη ώστε να δοθούν παραδείγματα επιθυμητής συμπεριφοράς.

Η συμπερίληψη του ιστορικού συνομιλιών είναι σημαντική όταν οι οδηγίες χρήστη αναφέρονται σε προηγούμενα μηνύματα. Στο παραπάνω παράδειγμα, η τελευταία ερώτηση του χρήστη "Πού παίχτηκε;" έχει νόημα μόνο στο πλαίσιο των προηγούμενων μηνυμάτων για το World Series του 2020. Επειδή τα μοντέλα δεν έχουν μνήμη προηγούμενων αιτημάτων, όλες οι σχετικές πληροφορίες πρέπει να παρέχονται ως μέρος του ιστορικού συνομιλιών σε κάθε αίτημα. Εάν μια συνομιλία δεν μπορεί να χωρέσει εντός του ορίου των tokens του μοντέλου, θα πρέπει να συντομευθεί με κάποιο τρόπο. Στο εν λόγω μοντέλο το όριο token είναι 4096.

Η απάντηση στην παραπάνω κλήση του API είναι της μορφής:

```
{
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "message": {
        "content": "The 2020 World Series was played in Texas at Globe Life Field in Arlington.",
        "role": "assistant"
      },
      "logprobs": null
    }
  ],
  "created": 1677664795,
  "id": "chatcmpl-7QyqpwdfhqwajicIEznoc6Q47XAYW",
  "model": "gpt-3.5-turbo-0613",
  "object": "chat.completion",
  "usage": {
    "completion_tokens": 17,
    "prompt_tokens": 57,
    "total_tokens": 74
  }
}
```

Και μπορεί να εξαχθεί ως εξής:

```
response ['choices'] [0] ['message'] ['content']
```

### 3.2.2 *Llama-2-chat*

Το Llama-2-chat-7b είναι ένα Μεγάλο Γλωσσικό Μοντέλο το οποίο αναπτύχθηκε και κυκλοφόρησε από την εταιρεία Meta. Ανήκει στην οικογένεια Llama-2 από LLMs, μια συλλογή από προ-εκπαιδευμένα και fine-tuned generative μοντέλα [27]. Ο αριθμός των παραμέτρων στα μοντέλα αυτά ποικίλει από 7 δισεκατομμύρια σε 70 δισεκατομμύρια. Το συγκεκριμένο που χρησιμοποιήθηκε έχει προσαρμοστεί για χρήση σε διαλογικές περιπτώσεις και έχει 7B παραμέτρους και 4096 tokens παράθυρο. Χρησιμοποιεί μια βελτιστοποιημένη έκδοση της αρχιτεκτονικής μετασηματιστή (transformer). Τέλος, το μοντέλο έχει εκπαιδευθεί κατά κύριο λόγο σε δεδομένα στην αγγλική γλώσσα. Η εκπαίδευσή του σε δεδομένα από άλλες γλώσσες είναι περιορισμένη. Ως εκ τούτου η επίδοσή του σε άλλη γλώσσα πλην της αγγλικής είναι χαμηλή.

### 3.2.3 *Mistral 7B*

Το open-source αυτό μοντέλο mistral-7b-instruct-v0.1.Q4\_0.gguf αναπτύχθηκε από την εταιρεία Mistral AI. Έχει 7.3 δισεκατομμύρια παραμέτρους και είναι η προσαρμοσμένη (fine-tuned) για διαλογική χρήση έκδοσή του Mistral 7b. Σχεδιάστηκε με σκοπό να εξισορροπήσει την απόδοση και το υπολογιστικό κόστος, καθιστώντας το πιο εφαρμόσιμο σε πραγματικά σενάρια χρήσης. Για την επίτευξη αυτού του στόχου, το μοντέλο χρησιμοποιεί δύο βασικούς μηχανισμούς. Ο πρώτος είναι ο Grouped-query Attention (GQA), που επιτρέπει ταχύτερους χρόνους απόκρισης σε σύγκριση με τον μηχανισμό attention. Ο δεύτερος ονομάζεται Sliding Window Attention (SWA) και δίνει στο Mistral τη δυνατότητα να επεξεργάζεται αποτελεσματικά μεγαλύτερες ακολουθίες κειμένου. [28-31].

### 3.2.4 *Mini orca*

Το μοντέλο orca-mini-3b-gguf2-q4\_0.gguf δημιουργήθηκε βασιζόμενο στο Llama μοντέλο ως μοντέλο βάσης (foundation model - LFM) και είναι κατάλληλο για χρήση με υλικό βασικών απαιτήσεων (entry-level hardware). Εκπαιδεύτηκε σε σύνολα δεδομένων τα οποία δημιουργήθηκαν βάσει των προσεγγίσεων και των περιορισμών που περιγράφονται στην εργασία “Orca: Progressive Learning from Complex Explanation Traces of GPT-4” [32,33]. Η βασική προσέγγιση είναι το explanation tuning, η τεχνική βελτιστοποίησης του μοντέλου στην οποία τα δεδομένα εκπαίδευσης (training data) εμπλουτίζονται με λεπτομερείς εξηγήσεις για την διαδικασία συλλογισμού βάσει της οποίας δίνεται κάθε απόκριση. Πέρα, δηλαδή, από την εκπαίδευση με ζευγάρια ερωτήσεων – απαντήσεων, στην περίπτωση του explanation tuning περιλαμβάνονται και τα βήματα ή η λογική που χρησιμοποιήθηκε για την απάντηση. Με τον τρόπο αυτό, το μοντέλο δεν μαθαίνει μόνο τις σωστές απαντήσεις αλλά και την υποκείμενη συλλογιστική διαδικασία, επιτρέποντάς του να κατανοήσει τη λογική και το σκεπτικό πίσω από διαφορετικές εργασίες.

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

### 3.2.5 GPT4All

Πέραν του **gpt-3.5-turbo** το οποίο είναι προσβάσιμο μέσω του OpenAI API και της χρήσης API key, για τα υπόλοιπα τρία μοντέλα χρησιμοποιήθηκε η εφαρμογή GPT4All [40].

Το GPT4All είναι ένα οικοσύστημα λογισμικού ανοιχτού κώδικα για την εκπαίδευση και την ανάπτυξη ισχυρών και προσαρμοσμένων Μεγάλων Γλωσσικών Μοντέλων που εκτελούνται τοπικά, σε CPU με απαιτήσεις επιπέδου καταναλωτή (consumer grade). Έτσι δίνεται η δυνατότητα να αξιοποιούνται LLMs ανοιχτού κώδικα και με τοπική χρήση, χωρίς να απαιτείται σύνδεση στο διαδίκτυο και χωρίς API κλειδί. Τα μοντέλα που υποστηρίζει έχουν αρχεία μεγέθους 3-8 GB και έχει βελτιστοποιηθεί, ώστε να επιτρέπει την απόκριση μοντέλων 3-13B παραμέτρων σε φορητούς υπολογιστές ή/και εξυπηρετητές. Τα υποστηριζόμενα μοντέλα έχουν εκπαιδευτεί με διαφορετικές αρχιτεκτονικές (πχ. LLaMA, GPT-J ) και κάποια από αυτά επιτρέπουν την εμπορική χρήση, ενώ αυτή απαγορεύεται σε άλλα.

Ένα από τα χαρακτηριστικά του GPT4All είναι το Server Mode. Αυτό σημαίνει ότι η GPT4All Chat εφαρμογή διαθέτει μια ενσωματωμένη λειτουργία διακομιστή API. Η λειτουργία επιτρέπει την αλληλεπίδραση με προγραμματιστικό τρόπο των υποστηριζόμενων LLMs μέσω του OpenAI API (χωρίς την απαίτηση της χρήσης API key).

```
import openai

openai.api_base = "http://localhost:4891/v1"
#openai.api_base = "https://api.openai.com/v1"

openai.api_key = "not needed for a local LLM"
```

Με τον τρόπο αυτό έγινε δυνατή η εξαγωγή των απαντήσεων από τα τρία μοντέλα.

## 3.3 Υλοποίηση

Καθώς σκοπός της εργασίας είναι να διερευνηθούν οι δυνατότητες των LLMs σχετικά με την εξαγωγή πληροφορίας από απλό (μη δομημένο) κείμενο, δεν έγινε επεξεργασία στις περιγραφές των αγγελιών εργασίας. Επιπλέον, η προσέγγιση zero-shot επιλέχθηκε σχετικά με την προτροπή. Η μέθοδος αυτή περιλαμβάνει τη δημιουργία προτροπών που θα επιτρέπουν στα LLM να εξάγουν πληροφορίες από τις περιγραφές θέσεων εργασίας χωρίς προηγούμενη εξειδικευμένη εκπαίδευση. Αυτή η προσέγγιση επιτρέπει την αξιολόγηση της έμφυτης ικανότητας των LLM να αναλύουν και να ερμηνεύουν σημασιολογικά άγνωστα δεδομένα κειμένου [70]. Οι προτροπές σχεδιάστηκαν για να κατευθύνουν τα LLM στην ανάλυση του περιεχομένου κειμένου και στη δόμηση των εξαγόμενων πληροφοριών σε μια συγκεκριμένη μορφή JSON. Αυτή η αξιολόγηση βοηθά στην εκτίμηση του πόσο αποτελεσματικά τα μοντέλα οργανώνουν και διαχειρίζονται δεδομένα σε πρακτικές εφαρμογές.

Στη φάση υλοποίησης, τα LLMs δόθηκε η εντολή να εντοπίσουν και να εξάγουν συγκεκριμένες πληροφορίες από τις αγγελίες θέσεων εργασίας, συμπεριλαμβανομένων του τίτλου της θέσης, της εταιρείας, τον κλάδο βιομηχανίας, της τοποθεσίας και των απαιτούμενων επαγγελματικών δεξιοτήτων. Αυτή η φάση αξιολόγησε την ικανότητα των μοντέλων να

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

εντοπίζουν και να οργανώνουν συστηματικά τα διάφορα συστατικά μέσα στο κείμενο της φυσικής γλώσσας.

Μετά τη διαδικασία εξαγωγής, τα ακατέργαστα αποτελέσματα μετατράπηκαν σε δομημένη μορφή CSV. Αυτός ο μετασχηματισμός ήταν απαραίτητος, καθώς εξασφάλιζε ότι τα δεδομένα από διαφορετικά LLM ήταν ενοποιημένα σε μια μορφή κατάλληλη για συγκριτική ανάλυση.

Για παράδειγμα, θεωρήστε μια αγγελία εργασίας ως εισαγωγή:

*“Elivi skiathos situated in Skiathos island and its aim is to create truly memorable 5-star hide away experiences. Inspired from the unique nature and heritage of the region, the renowned greek hospitality, our main aim is to deliver a more personalized and intuitive service, and to form tailored guest experiences. If the provision of discrete quality services in a luxury environment energizes, you, we will be glad to receive your resume. Front office receptionist we are looking to recruit an experienced FO receptionist responsible to provide customer focused service to the guests, ensuring their stay will become an unforgettable experience.*

*Key responsibilities: responsible for day-to-day operations of the front office, such as handling guest arrivals and departures, dealing in all FO related tasks and procedures. Operate relevant switchboard, call center and pms software as per all front office related activities identifies and ensures the highest standards of guest care is maintained, deals with any special request, complaints etc. Works closely with the housekeeping department for room turn around in order to allocate arriving guest in accommodation with a minimum of lost time. Must have a pleasant personality with excellent communication and interpersonal skills that promotes a helpful and professional image to the client and gives full cooperation to any customer requiring assistance. Requirements: 3-5 years working experience in a relevant position in 5 hotel hospitality degree written and verbal communication skills (english, greek, italian) proficiency in computer skills (windows, hotel pms and several industry platforms) team spirit and customers-oriented approach benefits: competitive remuneration package exciting and dynamic working environment food & accommodation (if not a local resident).”*

Για την δημιουργία της προτροπής δοκιμάστηκαν αρκετές διαφορετικές εκδοχές στο openai playground, μέχρις ότου να επιλεχθεί η τελική εκδοχή.

```
prompt_raw = ""Your task is to parse an unstructured job posting and turn it into a JSON containing the most important information. The job posting can describe one or more jobs at the same company. The JSON should consist of the following information:  
- id (field name: "id", field type: num)  
- The company name (field name: "companyName", field type: string)  
- The job title (field name: "jobTitle", field type: string),  
- the industry of the job(field:"industry:", field type:string)  
- the location of the job (field name: "location", field type: string  
- any specific soft skills that might be stated (field name: "softskills", field type: string).  
- any specific hard skills that might be stated (field name: "hardskills", field type: string).  
In general, if certain information is not stated, set the respective field to null.  
If the company seeks more than one person for the same role, include the role only once.
```



## Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

This is the job posting:

%s

The structured JSON representation is:

```
```json
{"job": [
  {
    "id": %s,
    "companyName": string,
    "jobTitle": string,
    "industry": string,
    "location": string,
    "softskills": string,
    "hardskills": string
  }
]
}""
```

Στην προτροπή δίνονται αναλυτικές οδηγίες στο μοντέλο για την εργασία που ζητείται να εκτελέσει, τα χαρακτηριστικά αυτών που ζητούνται, καθώς και η ακριβής δομή της JSON αναπαράστασης που είναι και τελικό ζητούμενο.

Η έξοδος JSON από το μοντέλο είναι της μορφής:

```
{"Job Title": " front office receptionist",
  "Company": " elivi skiathos",
  "Industry": " hospitality",
  "Location": " skiathos island",
  "Soft Skills": ['pleasant personality', 'excellent communication and interpersonal skills']",
  "Hard Skills": ['computer skills (windows, hotel pms and several industry platforms)']}
```

Εκτός από την προτροπή, δοκιμές έγιναν και για τις τιμές στις παραμέτρους των LLMs. Συγκεκριμένα, επιλέχθηκε η τιμή 0.2 για την θερμοκρασία, καθώς η συγκεκριμένη εργασία που ζητήθηκε από τα μοντέλα περιμένει ακριβείς απαντήσεις και όχι τυχαίες ή δημιουργικές εξόδους. Η Top p παράμετρος έμεινε στην default τιμή του 1, καθώς η σύσταση είναι για αλλαγή μίας από τις δύο παραμέτρους (θερμοκρασία ή Top P). Επίσης, επιλέχθηκε max\_length 500, αφού οι αναμενόμενες αποκρίσεις καλύπτονταν από την τιμή αυτή.

Καθώς διαπιστώθηκαν προβλήματα λόγω time-outs, τα αρχικά δεδομένα διασπάστηκαν σε μικρότερα αρχεία. Από το dataset διαβάζεται η στήλη Job Description που περιέχει το κείμενο της αγγελίας, όπως έχει συλλεχθεί από το διαδίκτυο. Τα αποτελέσματα είναι αρχεία csv με τις ζητούμενες απαντήσεις.

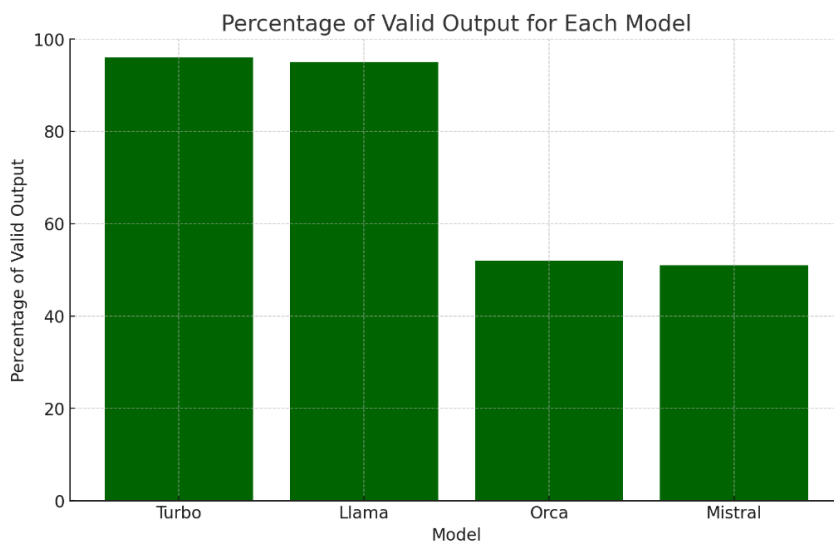
Αυτό το παράδειγμα αναδεικνύει την ικανότητα των LLMs να αναλύουν και να οργανώνουν τις αγγελίες εργασίας σε δομημένα δεδομένα, υπογραμμίζοντας τις δυνατότητές τους να οργανώνουν τις εργασίες εξαγωγής πληροφοριών για τους τελικούς χρήστες σε διάφορα υπολογιστικά περιβάλλοντα.

# 4

## Αποτελέσματα και Ανάλυση

### 4.1 Περιγραφή των αποτελεσμάτων

Η ανάλυση της απόδοσης των τεσσάρων LLM που χρησιμοποιήθηκαν οδήγησε σε ποικίλα αποτελέσματα, δείχνοντας διαφορές στην ακρίβεια των αποτελεσμάτων που παράγονται από κάθε μοντέλο. Η επεξεργασία ξεκίνησε με την αξιολόγηση της εγκυρότητας των εξόδων στα μοντέλα, όπως απεικονίζεται στο γράφημα [Σχήμα 18] παρακάτω. Η ανάλυση έδειξε ότι τα GPT 3.5 Turbo και Llama2 είχαν υψηλό ποσοστό έγκυρων εξόδων σε όλες σχεδόν τις καταχωρήσεις. Αντίθετα, τα μοντέλα Orca και Mistral παρουσίασαν έγκυρα αποτελέσματα για σημαντικά μικρότερο αριθμό.

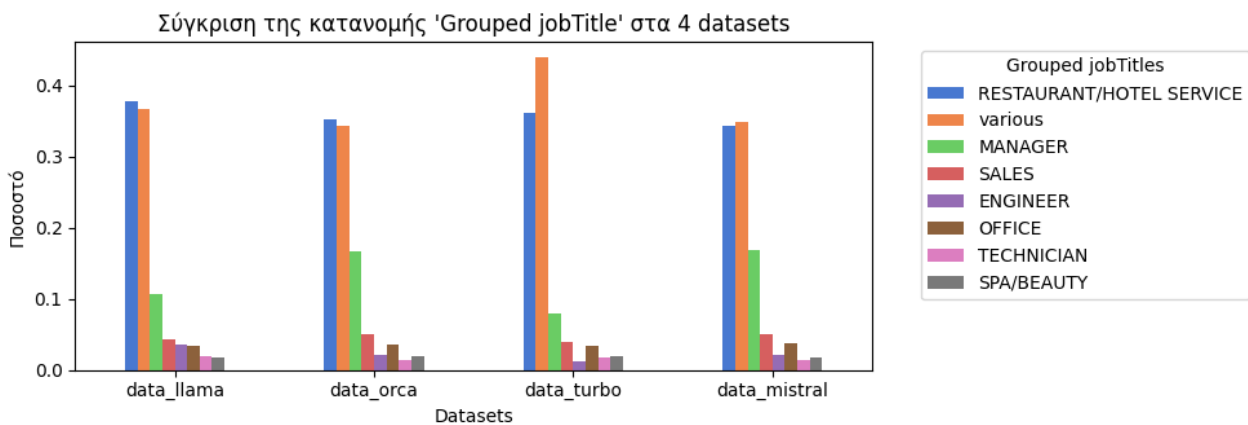


Σχήμα 18 Ποσοστό έγκυρων εξόδων ανά μεγάλο γλωσσικό μοντέλο

Βέβαια και στις περιπτώσεις που έχουμε έξοδο για την αγγελία από το μοντέλο, μπορεί να υπάρχει κενό στην απάντηση για κάποιο συγκεκριμένο ζητούμενο. Αυτό δεν είναι απαραίτητα λάθος του μοντέλου, καθώς μπορεί να μην υπάρχει η κατηγορία αυτή στην συγκεκριμένη αγγελία. Αυτό ισχύει κυρίως για τις οριζόντιες/ήπιες δεξιότητες (soft skills) και κάθετες/σκληρές δεξιότητες (hard skills - μη τεχνικές και τεχνικές δεξιότητες και γνώσεις).

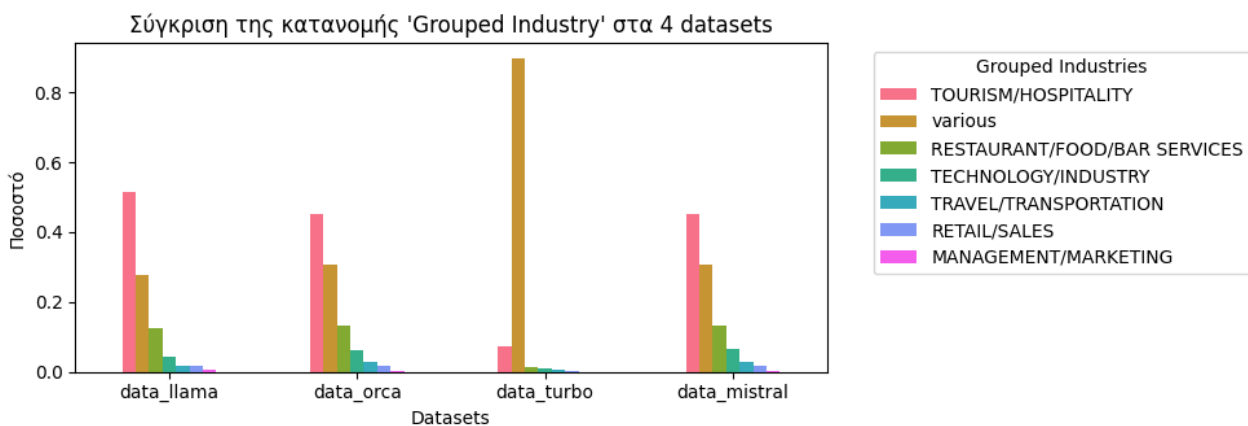
Ακολουθεί μια οπτική αποτύπωση των αποτελεσμάτων. Σημειώνεται ότι, για πιο συμπακνωμένη αναπαράσταση των τιμών, έγινε μια προσπάθεια ομαδοποίησης των αποτελεσμάτων π.χ., στο Industry κάτω από την ταμπέλα Tourism/Hospitality μπήκαν οι τιμές Tourism, Hospitality, Tourism and hospitality, hotel κτλ. Κάτω από την τιμή various είναι τα υπόλοιπα, αρκετά, διαφορετικά αποτελέσματα. Επισημαίνεται ότι η ομαδοποίηση δεν είναι εξαντλητική, αλλά ενδεικτική για την αποτύπωση των τάσεων που περιέχονταν στις αγγελίες.

#### Κατανομή για Ομαδοποιημένο Job Title (Τίτλος θέσης)



Σχήμα 19 Κατανομή για Ομαδοποιημένο Job Title (Τίτλος Θέσης)

#### Κατανομή για Ομαδοποιημένο Industry (κλάδος βιομηχανίας)



Σχήμα 20 Κατανομή για Ομαδοποιημένο Industry (κλάδος βιομηχανίας)



Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Word Cloud for Soft Skills in Turbo output



Σχήμα 23 Σύννεφο-λέξεων για τις ήπιες δεξιότητες που επέστρεψε το μοντέλο GPT3.5 Turbo

Word Cloud for Soft Skills in Mistral output



Σχήμα 24 Σύννεφο-λέξεων για τις ήπιες δεξιότητες που επέστρεψε το μοντέλο Mistral

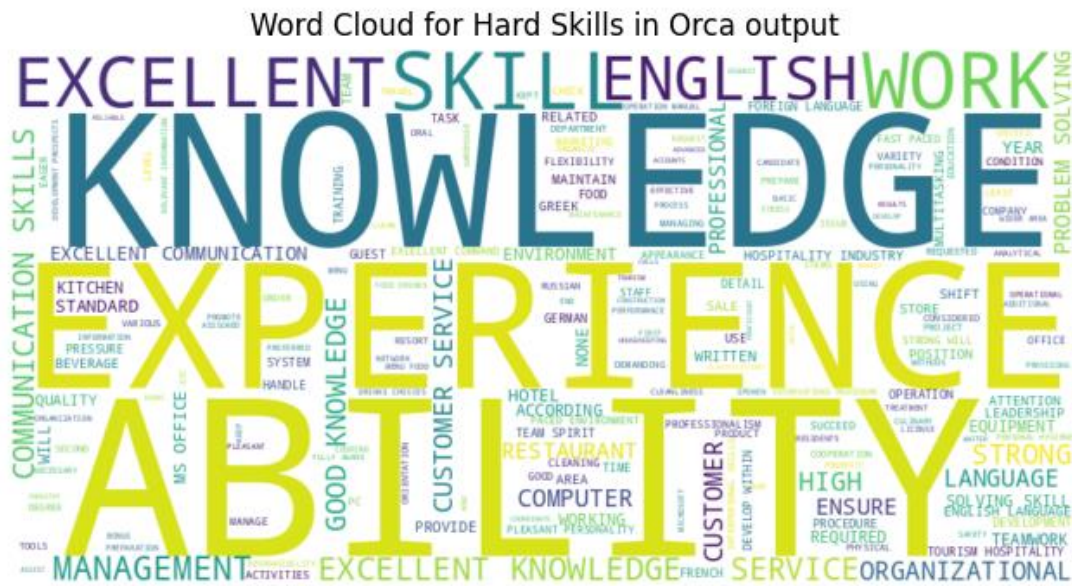








Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης



Σχήμα 29 Σύννεφο-λέξεων για τις σκληρές δεξιότητες που επέστρεψε το μοντέλο Orca

Στην κατηγορία αυτή υπάρχει μεγάλη ομοιότητα στις απαντήσεις στα μοντέλα, με την προηγούμενη εμπειρία και την γνώση αγγλικών να κυριαρχούν.

Από τα παραπάνω διαγράμματα διαπιστώνουμε ότι τα αποτελέσματα που έδωσαν τα μοντέλα συνάδουν και με το είδος των εργασιών που περιμέναμε (σχετικές με τουρισμό και συναφείς υπηρεσίες, κυρίως). Επίσης, διαπιστώνουμε ότι τα μοντέλα είναι γενικά κοντά μεταξύ τους στις απαντήσεις. Αναφορικά με τις πληροφορίες που εξήχθησαν για τις οριζόντιες/ήπιες δεξιότητες και για τις κάθετες/σκληρές δεξιότητες εκεί, τα μοντέλα ανταποκρίθηκαν λιγότερο καλά και έδωσαν απάντηση για λιγότερες εγγραφές.

Για πιο ασφαλή και συγκεκριμένα συμπεράσματα όμως, σχετικά με το πώς προσέγγισαν την χειρόκινητη εξαγωγή πληροφορίας τα μοντέλα, χρειάζεται να γίνει μια σύγκριση με τα δεδομένα που δόθηκαν στο αρχικό dataset.

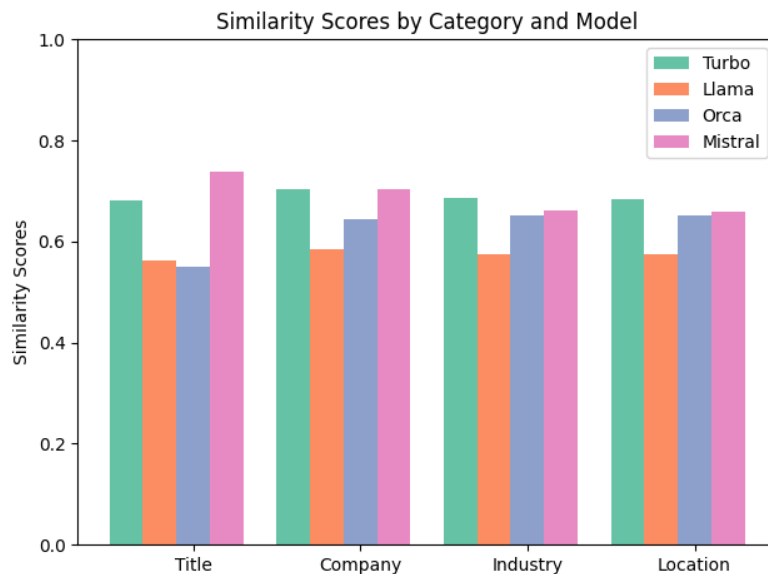
Παρακάτω εξηγείται η μέθοδος που ακολουθήθηκε για την σύγκριση αυτή:

Για την εξαγωγή συμπερασμάτων σχετικά με την ακρίβεια των αποτελεσμάτων, έγινε, με την χρήση του μοντέλου BERT, εξαγωγή των embeddings για τις στήλες Job Title, Company, Location και Industry. Κατόπιν, χρησιμοποιήθηκε η ομοιότητα συνημίτονου (cosine similarity) ως μέθοδος για τον υπολογισμό των αποτελεσμάτων σε σχέση με τα δοσμένα από το αρχικό dataset. Τα αποτελέσματα για την μέση τιμή του cosine similarity είναι ως εξής:

Model	Title	Company	Industry	Location
<b>Turbo</b>	0.6816	0.5620	0.5512	0.7390
<b>Llama</b>	0.7037	0.5853	0.6438	0.7044
<b>Orca</b>	0.6857	0.5748	0.6523	0.6608
<b>Mistral</b>	0.6848	0.5752	0.6524	0.6598

Σχήμα 30 Cosine Similarity





Σχήμα 31 Similarity Scores ανά μοντέλο

Η ανάλυση δεν φανέρωσε σημαντικές διακυμάνσεις στις βαθμολογίες ομοιότητας μεταξύ των μοντέλων, με τα GPT-3.5-Turbo και Mistral να παρουσιάζουν ελαφρώς καλύτερες επιδόσεις. Αναφορικά με τα αποτελέσματα ομοιότητας (Similarity scores) μπορούμε να παρατηρήσουμε τα εξής: το Gpt-3.5-Turbo δίνει σταθερά μεσαία προς υψηλά σκορ, υποδηλώνοντας ότι αυτό το μοντέλο συμπεριφέρεται σχετικά κοντά στην επιθυμητή συμπεριφορά, ότι είναι, δηλαδή, αρκετά αποτελεσματικό στην κατανόηση και την εξαγωγή των απαιτούμενων πληροφοριών. Το Llama2 δίνει σχετικά παρόμοιες βαθμολογίες με το Gpt-3.5-Turbo, χαμηλότερες βέβαια, καθώς και το ίδιο ποσοστό έγκυρων εξόδων. Το Mistral δείχνει καλές αποκρίσεις στις αγγελίες εργασίας για τις οποίες έδωσε απάντηση. Παρ' όλα αυτά, δυσκολεύεται, όπως το Orca, στην παραγωγή συνεπών αποκρίσεων.

Συγκεκριμένα, σε περιπτώσεις όπου η Orca και η Mistral απέτυχαν να δημιουργήσουν έγκυρα αποτελέσματα, οι απαντήσεις τους κατέληγαν σε μία προεπιλεγμένη (default) απόκριση που περιγράφει την αναμενόμενη δομή JSON, χωρίς να το συμπληρώνουν με εξαγόμενα δεδομένα. Αυτό το μοτίβο απόκρισης, με την χρήση ενός δομικού προτύπου αντί για συγκεκριμένα δεδομένα, υποδηλώνει πιθανά έναν εναλλακτικό μηχανισμό που χρησιμοποιείται από τα μοντέλα, όταν δυσκολεύονται να εξαγάγουν με ακρίβεια τις απαιτούμενες πληροφορίες.

Στις περιπτώσεις που δεν έδωσαν έγκυρο αποτέλεσμα, η απάντηση των μοντέλων ήταν, για παράδειγμα, της μορφής :

```
The unstructured job posting describes one or more jobs at the same company.
The JSON contains the following information:
- id (field name: "id", field type: num)
- The company name (field name: "companyName", field type: string)
- The job title (field name: "jobTitle", field type: string),
- the industry of the job(field:"industry:", field type:string)
- the location of the job (field name: "location", field type: string
- any specific soft skills that might be stated (field name: "softskills",
field type: string).
```

## Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

```
- any specific hard skills that might be stated (field name: "hardskills",  
field type: string).  
In general, if certain information is not stated, set the respective field to  
null.  
If the company seeks more than one person for the same role, include the role  
only once.
```

### Η

The unstructured job posting can be parsed into a JSON object as follows:

Ακόμα βέβαια και στις περιπτώσεις που δόθηκε έξοδος στην προκαθορισμένη μορφή, η έξοδος αυτή χρήζει βελτίωσης, ώστε να θεωρείται σωστή και να προσομοιάζει στη χειροκίνητη εξαγωγή πληροφορίας.

Γενικά, διαπιστώνουμε ότι η δομημένη εξαγωγή πληροφορίας από αδόμητο κείμενο είναι μια πρόκληση για τα LLMs, με τον τρόπο τουλάχιστον που επιχειρήθηκε στην συγκεκριμένη εργασία. Χρησιμοποιώντας το Gpt-3.5-Turbo ή το Llama2 μια πιο αποτελεσματική τελική απόκριση θα ήταν διασφαλισμένη. Καθώς όμως, η τεχνολογία στην οποία βασίζονται τα μοντέλα είναι σε διαρκή εξέλιξη, μια διαφορετική προσέγγιση θα μπορούσε να αποφέρει καλύτερα αποτελέσματα.

# 5

## *Συζήτηση και Μελλοντική Έρευνα*

Η εργασία ανέλυσε τη δυνατότητα των Μεγάλων Γλωσσικών Μοντέλων (LLMs) στην εξαγωγή δομημένων πληροφοριών από μη δομημένα κείμενα, καταγράφοντας τις διαφορετικές επιδόσεις τους. Αν και τα LLMs έχουν τις ικανότητες για να χρησιμοποιηθούν σε ευρύ φάσμα εφαρμογών, τα ευρήματα υποδηλώνουν ότι η αποτελεσματικότητά τους στην εξαγωγή δομημένων πληροφοριών ποικίλλει. Πιο συγκεκριμένα, τα μοντέλα GPT-3.5-Turbo και Llama2 ξεχώρισαν λόγω της υψηλότερης επίδοσής τους, ενώ τα Mistral και Orca έδωσαν σημαντικά λιγότερες έγκυρες εξόδους.

Οι παρατηρούμενες διαφορές στην απόδοση και τις μεθόδους διαχείρισης σφαλμάτων ανάμεσα στα μοντέλα παρέχουν ενδιαφέρουσες πληροφορίες για τον τρόπο λειτουργίας τους. Αυτό τονίζει την ανάγκη για πρόσθετη βελτιστοποίηση για την κάλυψη των πρακτικών απαιτήσεων των τελικών χρηστών. Λόγω των ραγδαίων εξελίξεων στον τομέα, αξίζει να ληφθεί υπόψη ότι τα εν λόγω εμπόδια θα μπορούσαν να ξεπεραστούν με την διερεύνηση καινοτόμων προσεγγίσεων ή προόδων στην τεχνολογία των μεγάλων γλωσσικών μοντέλων.

Η μελλοντική έρευνα θα πρέπει να λάβει υπόψη την εξερεύνηση αναδυόμενων μοντέλων και καινοτόμων μεθοδολογιών εκπαίδευσης αυτών, προκειμένου να ενισχυθούν οι δυνατότητες δομημένης εξαγωγής πληροφοριών. Μια ολοκληρωμένη ανάλυση των κατηγοριών σφαλμάτων και της επίδρασης της πολυπλοκότητας των δεδομένων εισόδου στην απόδοση του μοντέλου θα μπορούσε να προσφέρει στοχευμένες προσεγγίσεις για την βελτίωση της ακρίβειας. Καθώς η επεξεργασία φυσικής γλώσσας και η μηχανική εκμάθηση συνεχίζουν να εξελίσσονται, υπάρχουν συνεχείς ευκαιρίες για καινοτομία.

Με βάση τα συμπεράσματα αυτής της εργασίας, είναι επιτακτική ανάγκη να αναγνωριστούν οι ευρύτερες κοινωνικοοικονομικές επιπτώσεις της χρήσης LLM για δομημένη εξαγωγή πληροφοριών από αγγελίες εργασίας. Η δυνατότητα βελτίωσης των διαδικασιών πρόσληψης,

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

προσαρμογής των εκπαιδευτικών προγραμμάτων στις ανάγκες της αγοράς και ενίσχυσης της χάραξης πολιτικής με γνώσεις βάσει δεδομένων, ανοίγει νέους δρόμους για την κοινωνική πρόοδο. Αυτό υπογραμμίζει τη σημασία των ηθικών κριτηρίων, του απορρήτου των δεδομένων και του μετριασμού των αλγοριθμικών προκαταλήψεων (biases) στην μελλοντική ανάπτυξη και εφαρμογή LLM, διασφαλίζοντας ότι τα οφέλη της τεχνητής νοημοσύνης στην ανάλυση της αγοράς εργασίας θα κατανοούνται δίκαια σε όλους τους τομείς της κοινωνίας.

## 5.1 Προτάσεις για επέκταση της εργασίας

Με βάση τις τεχνολογικές εξελίξεις, όπως παρουσιάζονται το τελευταίο χρονικό διάστημα, δίνονται παρακάτω κάποιες προτάσεις για μελλοντική βελτίωση/έρευνα στο συγκεκριμένο πεδίο:

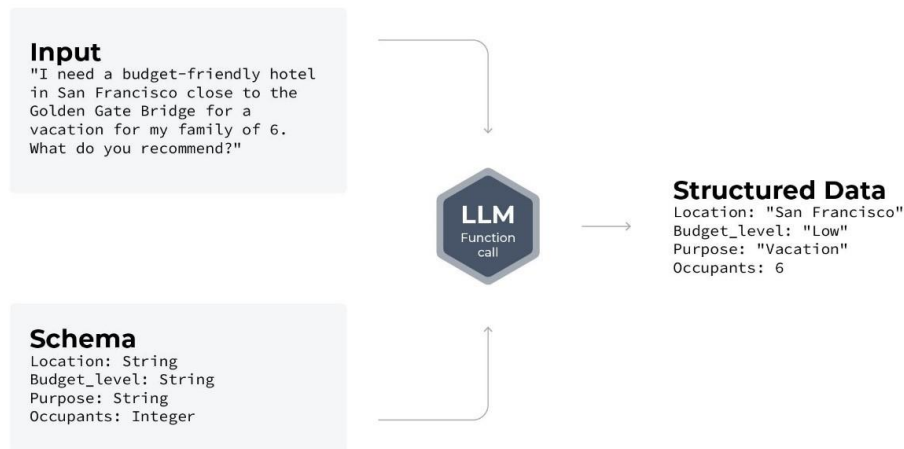
- **Χρήση μοντέλων με περισσότερες παραμέτρους.** Οι συγκεκριμένες δοκιμές έγιναν σε μοντέλα με 7b όπως το Llama-2 ή το Mistral . Θα μπορούσαν να χρησιμοποιηθούν οι 13b εκδοχές τους για σύγκριση. Είναι πιθανό οι περισσότερες παράμετροι να βελτιώνουν την zero-shot απόδοση των μοντέλων.
- **Δοκιμή διαφορετικών προτροπών.** Παρόλο που για την επιλογή της τελικής προτροπής είχαν προηγηθεί αρκετές δοκιμές με διαφορετικές εκδοχές, διαφορετικές τεχνικές prompting όπως few-shot ή chain-of-thought θα μπορούσαν να εξερευνηθούν.
- Προσπάθεια κατεύθυνσης της εξόδου στην επιθυμητή μορφή με την χρήση της **βιβλιοθήκης Kor** [52]. Η Kor είναι μια βιβλιοθήκη της Python με ενσωματωμένη υποστήριξη για την εξαγωγή δομημένων δεδομένων από κείμενο. Ουσιαστικά, προσθέτει επιπλέον λειτουργικότητα πάνω από το LLM για να βοηθήσει στην εξαγωγή.

Χαρακτηριστικά της Kor, σύμφωνα με τον δημιουργό της:

- Υποστήριξη για τον ορισμό σχήματος δεδομένων (data scheme) με ορισμούς τύπων, περιγραφές πεδίων και παραδειγμάτων εξαγωγής.
- Παρέχει ενσωματωμένη προτροπή εξαγωγής. Υποστηρίζει επίσης την προσθήκη custom prompts.
- Χρήση CSV ή JSON κωδικοποιητών δεδομένων
- Επικύρωση εξαγόμενων δεδομένων σύμφωνα με το σχήμα (scheme)
- Διαχείριση λαθών εξαγωγής

Η Kor φαίνεται μια προσπάθεια για την καθοδήγηση της εξόδου των μοντέλων σε συγκεκριμένες μορφές. Δεν έχει δοκιμαστεί στο συγκεκριμένο πρόβλημα, αλλά δείχνει να έχει πιθανότητες βελτίωσης των αποτελεσμάτων. Παραμένει βέβαια μια νέα, εν εξελίξει προσπάθεια με πιθανές αστοχίες και λάθη.

- **Function calling.** Το τελευταίο διάστημα έχει προστεθεί η δυνατότητα του ορισμού συναρτήσεων από τον χρήστη σε κάποια LLMs. Στο παρακάτω διάγραμμα βλέπουμε σχηματικά τον ορισμό συναρτήσεων:



Σχήμα 32. Function Calling [53]

Στην τεχνική αυτή ο χρήστης ορίζει την συνάρτηση, η οποία και περιέχει συγκεκριμένο scheme (σχήμα) για το πώς θα είναι τα δεδομένα στην επιθυμητή έξοδο. Η συνάρτηση είναι όρισμα στο Chat.completion.request (σημείωση: η επιλογή συνάρτησης για τον καθορισμό της εξόδου του μοντέλου δεν αποκλείει το μοντέλο να έχει «παραισθήσεις» (hallucinations και να παράγει ψευδά αποτελέσματα) Επιπλέον, οι συναρτήσεις μετράνε ως tokens στην κλήση του μοντέλου άρα και στο όριο των tokens αυτού.

Χρησιμοποιώντας το LangChain, ένα open-source framework που βοηθάει στην δημιουργία εφαρμογών για LLMs, είναι δυνατή η παράλληλη κλήση συναρτήσεων, μειώνοντας έτσι τις κλήσεις προς το LLM άρα και εξοικονομώντας χρόνο.

- **Finetuning** των μοντέλων με αντίστοιχα δεδομένα.

Το Finetuning είναι μια τεχνική για την εκπαίδευση των μοντέλων σε συγκεκριμένα δεδομένα (πχ. ιατρικά δεδομένα) . Έτσι το μοντέλο ειδικεύεται για την χρήση του σε συγκεκριμένα αντικείμενα. Στην πράξη το fine tuning είναι η διαδικασία της προσαρμογής των παραμέτρων ενός LLM για μια συγκεκριμένη εργασία. Για να γίνει αυτό θα πρέπει να έχουμε δεδομένα εκπαίδευσης Στο πλήρες finetuning γίνεται βελτιστοποίηση των βαρών σε όλα τα επίπεδα του μοντέλου. Αυτό απαιτεί χρόνο και υπολογιστική ισχύ. Σε αρκετές περιπτώσεις καλύτερα και σίγουρα πιο γρήγορα αποτελέσματα έχει το finetuning με την χρήση της μεθόδου LoRa ή της παραλλαγής της QLoRa. Στις μεθόδους αυτές μειώνεται ο αριθμός των παραμέτρων που βελτιστοποιούνται κατά το finetuning στο νευρωνικό δίκτυο, με αποτέλεσμα να είναι πιο γρήγορη και λιγότερη απαιτητική υπολογιστικά η βελτιστοποίηση τους.

- **Μοντέλο GoLLIE.**

Εξαγωγή πληροφοριών από αγγελίες εργασίας με χρήση εργαλείων από διαλογικά συστήματα τεχνητής νοημοσύνης

Τέλος, στο πλαίσιο του finetuning, ενδιαφέρον παρουσιάζει το μοντέλο GoLLIE. Στο μοντέλο αυτό έχει γίνει fine tuning με σκοπό την βελτίωση της απόδοσης του baseline μοντέλου (χρησιμοποίησαν το Code-LLaMA) στο να ακολουθεί οδηγίες επισημείωσης (Annotation guidelines improvement) και σύμφωνα με την δημοσίευση [49] βελτιώνει την zero-shot εξαγωγή πληροφορίας.

## Βιβλιογραφία

1. P. Kulkarni, A. Mahabaleshwarkar, M. Kulkarni, N. Sirsikar and K. Gadgil, "Conversational AI: An Overview of Methodologies, Applications & Future Scope," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-7, doi: 10.1109/ICCUBEA47591.2019.9129347.
2. Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana "Practical Natural Language Processing\_ A Comprehensive Guide to Building Real-World NLP Systems" -O'Reilly Media, Inc. (2020)
3. "What Is Natural Language Processing?" IBM <https://www.ibm.com/topics/natural-language-processing>.
4. Alammar, Jay. "The Illustrated Transformer." The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. <https://jalammar.github.io/illustrated-transformer>.
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need". In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>
6. Harishdatab. "Unveiling the Power of Large Language Models (Llms)." Medium (2023). <https://medium.com/@harishdatab/unveiling-the-power-of-large-language-models-llms-e235c4eba8a9>
7. Amanatullah. "Transformer Architecture Explained." Medium, (2023) <https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c>.
8. Alammar, Jay. "The Illustrated Word2vec." The Illustrated Word2vec – Jay Alammar – Visualizing machine learning one concept at a time. <http://jalammar.github.io/illustrated-word2vec/>.
9. "What Is a Large Language Model?: A Comprehensive Llms Guide." Elastic. <https://www.elastic.co/what-is/large-language-models>.
10. Kroon, A., Welbers, K., Trilling, D., & van Atteveldt, W. (2023). "Advancing Automated Content Analysis for a New Era of Media Effects Research: The Key Role of Transfer Learning." Communication Methods and Measures, 1-21. D OI: 10.1080/19312458.2023.2261372
11. "Fine-tuning - OpenAI APP". <https://platform.openai.com/docs/guides/fine-tuning>

12. Alammar, Jay. “*The Illustrated GPT-2 (Visualizing Transformer Language Models)*.” The Illustrated GPT-2 (Visualizing Transformer Language Models) – Jay Alammar – Visualizing machine learning one concept at a time. <http://jalammar.github.io/illustrated-gpt2/>.
13. Devlin, Jacob, et al. “*Bert: Pre-training of deep bidirectional transformers for language understanding*” <https://doi.org/10.48550/arXiv.1810.04805> (2018).
14. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “*Language models are unsupervised multitask learners,*” OpenAI blog, vol. 1, no. 8, p. 9, 2019.
15. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal et al. “*Language models are few-shot learners*”. In Advances in NeurIPS, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. <https://doi.org/10.48550/arXiv.2005.14165>
16. Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). “*Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*”. arXiv preprint arXiv:2310.14735.
17. “*Best practices for prompt engineering with openai API*” | openai help center. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api>
18. “*Prompt engineering - openai API*”. <https://platform.openai.com/docs/guides/prompt-engineering>
19. Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester et al “*Finetuned Language Models Are Zero-Shot Learners*” (2021) <https://doi.org/10.48550/arXiv.2109.01652>
20. Nextra “*Prompt Engineering Guide*” <https://www.promptingguide.ai/>
21. Reynolds, Laria, and Kyle McDonell. “*Prompt programming for large language models: Beyond the few-shot paradigm.*” Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.
22. Panagiotis Skondras “*Panagiotis-Skondras/Informatics: MDPI Information Paper*”, GitHub. <https://github.com/Panagiotis-Skondras/informatics>
23. D. Vukadin, A. S. Kurdija, G. Delač and M. Šilić, “*Information Extraction From Free-Form CV Documents in Multiple Languages,*” in IEEE Access, vol. 9, pp. 84559-84575, (2021), doi: 10.1109/ACCESS.2021.3087913.
24. Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, et al. (2022). “*Emergent abilities of large language models.*” doi:10.48550/arXiv.2206.07682
25. J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y et al. “*A comprehensive capability analysis of gpt-3 and gpt-3.5 series models*” (2023) doi:10.48550/arXiv.2303.10420
26. “*OpenAI developer platform*”. <https://platform.openai.com/docs/models/gpt-3-5>
27. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. “*Llama 2: Open foundation and fine-tuned chat models.*” (2023) doi: 10.48550/arXiv.2307.09288
28. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D.L., et al, (2023) “*Mistral 7B*” doi:10.48550/arXiv.2310.06825
29. Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Shanghai. “*Gqa: Training generalized multi-query transformer models from multi-head checkpoints*”. (2023) doi:10.48550/arXiv.2305.13245.



30. Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. “*Generating long sequences with sparse transformers*” (2019) doi:10.48550/arXiv.1904.10509
31. Iz Beltagy, Matthew E Peters, and Arman Cohan. “*Longformer: The long-document transformer*” (2020) doi:10.48550/arXiv.2004.05150.
32. pankajmathur/orca\_mini\_3b – Hugging Face [https://huggingface.co/pankajma-thur/orca\\_mini\\_3b](https://huggingface.co/pankajma-thur/orca_mini_3b)
33. S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi and A. Awadallah. “*Orca: Progressive learning from complex explanation traces of GPT-4*”. (2023) doi:10.48550/arXiv.2306.02707
34. Mohammed (2023) “*Types of open source & closed source LLMs (Large language Models)*”, Medium. <https://medium.com/@techlatest.net/types-of-open-source-llms-large-language-models-3b7d8b8d1af2>
35. Yu, Hao, et al. “*Open, Closed, or Small Language Models for Text Classification?*” (2023). doi:10.48550/arXiv.2308.10092
36. Singh, Sonit. “*Natural language processing for information extraction.*” (2018) doi:10.48550/arXiv.1807.02383
37. Li, Nan, Bo Kang, and Tijn De Bie. “*LLM4Jobs: Unsupervised occupation extraction and standardization leveraging Large Language Model.*” (2023) doi:10.48550/arXiv.2309.09708
38. Jeong, Cheonsu. “*Fine-tuning and Utilization Methods of Domain-specific LLMs.*” (2024) doi:10.48550/arXiv.2401.02981.
39. Ghosh, Preetam, and Vaishali Sadaphal. “*JobRecoGPT--Explainable job recommendations using LLMs.*” (2023) doi:10.48550/arXiv.2309.11805
40. GPT4All <https://gpt4all.io>
41. API reference - openai API. <https://platform.openai.com/docs/api-reference>
42. LeCun, Y., Bengio, Y., & Hinton, G. (2015). “*Deep learning*”. *nature*, 521(7553), 436-444. doi:10.1038/nature14539
43. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). “*Improving language understanding by generative pre-training.*” <https://openai.com/research/language-unsupervised>
44. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, et al (2019). “*A guide to deep learning in healthcare.*” *Nature medicine*, 25(1), 24-29. doi: 10.1038/s41591-018-0316-z
45. Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018, February). “*Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction.*” In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 261-269). doi:10.48550/arXiv.1712.02136
46. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). “*On the dangers of stochastic parrots: Can language models be too big?*” In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). doi:10.1145/3442188.3445922
47. Howard, J., & Ruder, S. (2018). “*Universal language model fine-tuning for text classification.*” doi:10.48550/arXiv.1801.06146
48. Rajkomar, A., Oren, E., Chen, K. et al. “*Scalable and accurate deep learning with electronic health records.*” *npj Digital Med* 1, 18 (2018). doi:10.1038/s41746-018-0029-1
49. Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O. L., Rigau, G., & Agirre, E. (2023). “*Gollie: Annotation guidelines improve zero-shot information-extraction*”. arXiv preprint arXiv:2310.03668.

50. Eloundou, Tyna, et al. “*Gpts are gpts: An early look at the labor market impact potential of large language models.*” (2023). doi:10.48550/arXiv.2303.10130
51. Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. “*Chain-of-thought prompting elicits reasoning in large language models.*” *Advances in Neural Information Processing Systems* 35 (2022): 24824-24837.
52. Kor 1.0.1. <https://eyurtsev.github.io/kor/index.html>
53. “*Prepare for (better) structured data extraction*” DataTalks.Club. <https://datatalks.club/blog/prepare-for-better-structured-data-extraction.html>
54. Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama et al. “*Emergent abilities of large language models.*” arXiv preprint arXiv:2206.07682 (2022).
55. Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. “*Is chatgpt a general-purpose natural language processing task solver?.*” arXiv preprint arXiv:2302.06476 (2023).
56. Hadi, Muhammad Usman, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. “*Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects.*” *Authorea Preprints* (2023).
57. Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. “*A comprehensive overview of large language models.*” arXiv preprint arXiv:2307.06435 (2023).
58. Gilson, A. et al. “*How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment.*” *JMIR Med Educ* 9, e45312 (2023).
59. Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. “*Large language models in medicine.*” *Nature medicine* 29, no. 8 (2023): 1930-1940.
60. Al Nazi, Zabir, and Wei Peng. “*Large language models in healthcare and medical domain: A review.*” arXiv e-prints (2023): arXiv-2401.
61. Zhao, Huaqin, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu et al. “*Revolutionizing finance with llms: An overview of applications and insights.*” arXiv preprint arXiv:2401.11641 (2024).
62. Li, Yinheng, Shaofei Wang, Han Ding, and Hang Chen. “*Large language models in finance: A survey.*” In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374-382. 2023.
63. Deng, Li, and Dong Yu. “*Deep learning: methods and applications.*” *Foundations and trends in signal processing* 7.3–4 (2014): 197-387. <http://dx.doi.org/10.1561/20000000039>
64. Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, et al. “*Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond.*” (2023) <https://doi.org/10.48550/arXiv.2304.13712>
65. Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, et al. “*A survey of large language models*” (2023) <https://doi.org/10.48550/arXiv.2303.18223>

66. Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, et al. “Scaling language models: Methods, analysis & insights from training gopher.” (2021) <https://doi.org/10.48550/arXiv.2112.11446>
67. Liu, Xiao, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, et al. “Agentbench: Evaluating llms as agents.” (2023) <https://doi.org/10.48550/arXiv.2308.03688>
68. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, et al. “Recent advances in natural language processing via large pre-trained language models: A survey.” *ACM Computing Surveys*, 56(2), pp.1-40. (2023) <https://doi.org/10.1145/3605943>
69. Kaddour, Jean, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. “Challenges and applications of large language models.” (2023) <https://doi.org/10.48550/arXiv.2307.10169>
70. Wang, Zhiqiang, Yiran Pang, and Yanbin Lin “Large Language Models Are Zero-Shot Text Classifiers.” (2023) <https://doi.org/10.48550/arXiv.2312.01044>.
71. Wang, Yuqing, Yun Zhao, and Linda Petzold. “Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding.” (2023) <https://doi.org/10.48550/arXiv.2304.05368>
72. Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. “Can ChatGPT understand too? a comparative study on ChatGPT and fine-tuned BERT”. (2023) <https://doi.org/10.48550/arXiv.2302.10198>
73. Guha, Neel, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters et al. “Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.” *Advances in Neural Information Processing Systems* 36 (2024).
74. Cui, Jiayi, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. “Chatlaw: Open-source legal large language model with integrated external knowledge bases.” *arXiv preprint arXiv:2306.16092* (2023).
75. Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. “A survey on evaluation of large language models.” *ACM Transactions on Intelligent Systems and Technology* (2023).
76. GPT-3 (2024) Wikipedia. <https://en.wikipedia.org/wiki/GPT-3>.
77. Qader, Wisam A., Musa M. Ameen, and Bilal I. Ahmed. “An overview of bag of words; importance, implementation, applications, and challenges.” In *2019 International Engineering Conference (IEC)*, pp. 200-204. IEEE, 2019.
78. Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. “Text classification algorithms: A survey.” *Information* 10, no. 4 (2019): 150.
79. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space.” *arXiv preprint arXiv:1301.3781* (2013).
80. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. “Glove: Global vectors for word representation.” In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.

81. Lee, David. “*The Transformer Model: Revolutionizing Natural Language Processing.*” Medium, <https://medium.com/@lordmoma/the-transformer-model-revolutionizing-natural-language-processing-a16be54ddb1e>.
82. Papagiannopoulou, Eirini, and Grigorios Tsoumakas. “*A review of keyphrase extraction.*” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, no. 2 (2020): e1339.
83. Wang, Shuhe, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. “*GPT-NER: Named entity recognition via large language models.*” arXiv preprint arXiv:2304.10428 (2023).
84. Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li. “*A survey on deep learning for named entity recognition.*” IEEE transactions on knowledge and data engineering 34, no. 1 (2020): 50-70.
85. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). “*Deep contextualized word representations*” arXiv preprint arXiv:1802.05365
86. Yuan, Zhihang, Yuzhang Shang, Yang Zhou, Zhen Dong, Chenhao Xue, Bingzhe Wu, Zhikai Li et al. “*LLM Inference Unveiled: Survey and Roofline Model Insights.*” arXiv preprint arXiv:2402.16363 (2024).
87. McKenna, Nick, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. “*Sources of hallucination by large language models on inference tasks.*” arXiv preprint arXiv:2305.14552 (2023).
88. Dogra, Varun, Sahil Verma, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. “*A complete process of text classification system using state-of-the-art NLP models.*” Computational Intelligence and Neuroscience 2022 (2022).
89. Manikandan, Ankur. “*How Does Temperature Impact next Token Prediction in LLMS?*” Medium, <https://towardsdatascience.com/how-does-temperature-impact-next-token-prediction-in-llms-779bd908f2cf>.