# Σχολή Ανθρωπιστικών Επιστημών

## Ψηφιακές Ανθρωπιστικές Επιστήμες:

## μέθοδοι, εργαλεία, πρακτικές (ΨΑΕ)

Διπλωματική Εργασία:

«Ανίχνευση ειρωνείας σε ελληνόγλωσσα διαδικτυακά κείμενα: Συγκριτική αξιολόγηση υπολογιστικών προσεγγίσεων στην επεξεργασία φυσικής γλώσσας»

Όνομα Συγγραφέα: Άγγελος Ανδρεοσόπουλος

Επιβλέπων Καθηγητής: Διονύσιος Γούτσος

Πάτρα, Ιούνιος 2026

# Περίληψη

Η παρούσα εργασία πραγματεύεται την αυτόματη ανίχνευση ειρωνείας σε ελληνόγλωσσες διαδικτυακές κριτικές καταναλωτών μέσω συγκριτικής αξιολόγησης τριών υπολογιστικών προσεγγίσεων: ενός αλγορίθμου της Python που βασίζεται σε κανόνες (rule-based) και δύο Μεγάλων Γλωσσικών Μοντέλων (ΜΓΜ, LLMs), του Claude (Anthropic) και του DeepSeek (DeepSeek AI). Η αξιολόγηση πραγματοποιείται έναντι ανθρώπινης επισημείωσης (πραγματική κατάσταση, ground truth) σε ένα σώμα 500 κριτικών καταναλωτών από διαδικτυακές πλατφόρμες. Τα αποτελέσματα αναδεικνύουν τον τρόπο με τον οποία τα μοντέλα ανιχνεύουν ή όχι την ειρωνεία, τα σημεία που τα δυσκολεύουν και τα ποσοστά συμφωνίας μεταξύ τους. Τα ευρήματα αναδεικνύουν ένα κεντρικό και απροσδόκητο αποτέλεσμα: ο αλγόριθμος που είναι βασισμενος σε ορισμένους κανόνες (rule-based) υπερτερεί και των δύο ΜΓΜ, επιτυγχάνοντας F1 = 0.686 και Δείκτη συμφωνίας κ = 0.495, υποδηλώνοντας ότι τα λεξιλογικά σήματα που ανιχνεύει αντιστοιχούν επαρκώς στα μοτίβα ειρωνείας του συγκεκριμένου σώματος κειμένων. Το Claude εμφανίζει υψηλά Ανάκληση (Recall: 0.608), αλλά συστηματική υπεργενίκευση με 141 ψευδώς θετικές ταξινομήσεις (Ακρίβεια, Precision = 0.456), ενώ το DeepSeek κινείται στον αντίποδα με εξαιρετικά υψηλή Ακρίβεια (0.857) αλλά εξαιρετικά χαμηλή Ανάκληση (0.216), με αποτέλεσμα να μην αναγνωρίζει την πλειονότητα των ειρωνικών κειμένων. Ιδιαίτερα αποκαλυπτική είναι η σχεδόν μηδενική συμφωνία μεταξύ των δύο ΜΓΜ (κ = 0.044), που υποδηλώνει ότι τα μοντέλα δεν προσεγγίζουν το ίδιο εννοιολογικό φαινόμενο, αλλά ανιχνεύουν ποιοτικά διαφορετικές εκδοχές ειρωνείας

Τα ευρήματα συζητούνται υπό το πρίσμα των ιδιαιτεροτήτων της ελληνικής ως γλώσσας χαμηλών πόρων, της ευαισθησίας των ΜΓΜ στη διατύπωση των προτροπών (prompts) και των περιορισμών της μονοαξιολογητικής επισημείωσης. Ως προς τις προοπτικές, η εργασία αναδεικνύει τρεις κρίσιμες κατευθύνσεις: την ανάπτυξη εξειδικευμένου ελληνόγλωσσου σώματος κειμένων ειρωνείας με πολλαπλή επισήμανση, την εξειδικευμένη εκπαίδευση (fine-tuning) πολυγλωσσικών μοντέλων σε ελληνικά δεδομένα, και την ανάπτυξη υβριδικών προσεγγίσεων που αξιοποιούν τα συμπληρωματικά πλεονεκτήματα των τριών συστημάτων. Η εργασία συνεισφέρει στην υπολογιστική επεξεργασία της ελληνικής γλώσσας, αναδεικνύοντας ότι η ανίχνευση ειρωνείας παραμένει ζήτημα προς επίλυση που απαιτεί όχι μόνο καλύτερα μοντέλα, αλλά και καλύτερους γλωσσικούς πόρους.

**ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ**

αλγόριθμος που βασίζεται σε κανόνες, γλώσσες χαμηλών πόρων, ειρωνεία, επεξεργασία φυσικής γλώσσας, Μεγάλα Γλωσσικά Μοντέλα

# Abstract

The thesis deals with the automatic detection of irony in Greek-language online consumer reviews through a comparative evaluation of three computational approaches: a rule-based Python algorithm and two Large Language Models (LLMs), Claude (Anthropic) and DeepSeek (DeepSeek AI). The evaluation is conducted against human annotations (ground truth) on a corpus of 500 consumer reviews from online platforms. The results highlight how the models do or do not detect irony, the areas where they struggle and the agreement rates among them. The findings reveal a significant and unexpected result: the rule-based algorithm outperforms both LLMs, achieving an F1 score of 0.686 and a Cohen's k of 0.495, suggesting that the lexical cues it detects adequately correspond to the irony patterns in this specific corpus. Claude exhibits high Recall (0.608), but systematic overgeneralization with 141 false positives (Precision = 0.456), while DeepSeek is at the opposite end of the spectrum with extremely high Precision (0.857), but extremely low Recall (0.216), failing to identify most ironic texts. Particularly revealing is the near-zero agreement between the two LLMs ($\kappa = 0.044$), suggesting that the models do not address the same conceptual phenomenon, but detect qualitatively different forms of irony.

The findings are discussed considering the particularities of Greek as a low-resource language, the sensitivity of LLMs to the wording of prompts and the limitations of single-label annotation. Regarding future directions, the paper highlights three critical areas: the development of a specialized Greek language corpus of irony with multi-label annotation, the fine-tuning of multilingual models on Greek data and the development of hybrid approaches that leverage the complementary strengths of the three systems. This work contributes to the computational processing of the Greek language and highlights that irony detection remains an open problem, requiring not only better models, but also better language resources.

**KEYWORDS**

irony detection, large language models, low-resource languages, natural language processing, rule-based algorithm

# Περιεχόμενα

# Κατάλογος Εικόνων



# Κατάλογος Πινάκων



# Ακρωνύμια – Συντομεύσεις

| | |
|---|---|
| ΕΦΓ | Επεξεργασία Φυσικής Γλώσσας |
| ΜΓΜ | Μεγάλα Γλωσσικά Μοντέλα |
| NLP | Natural Language Processing |

# 1. Εισαγωγή

Το πρώτο κεφάλαιο αποτελεί την εισαγωγή στην παρούσα εργασία και τοποθετεί το ερευνητικό πρόβλημα στο κατάλληλο επιστημονικό και κοινωνικό πλαίσιο. Αρχικά, προσδιορίζεται το αντικείμενο μελέτης και παρουσιάζεται το σώμα της έρευνας (1.1), ενώ στη συνέχεια αναλύεται γιατί η αυτόματη αναγνώριση ειρωνείας στις διαδικτυακές κριτικές αποτελεί πρόβλημα με ουσιαστικές πρακτικές συνέπειες (1.2). Ακολουθεί παρουσίαση των κυριότερων τομέων εφαρμογής της ανίχνευσης ειρωνείας (1.3) και διατύπωση του ερευνητικού σκοπού και των ερευνητικών ερωτημάτων που διέπουν τη μελέτη (1.4). Το κεφάλαιο ολοκληρώνεται με τα αναμενόμενα αποτελέσματα (1.5) και με μια συστηματική επισκόπηση των προκλήσεων που αντιμετωπίζουν τα αυτόματα συστήματα κατά την ανάλυση ελληνόγλωσσου κειμένου (1.6).

## 1.1. Το πρόβλημα της ειρωνείας

Η ικανότητα κατανόησης της γλώσσας αποτελεί ένα από τα θεμελιώδη προβλήματα της υπολογιστικής επεξεργασίας φυσικής γλώσσας (ΕΦΓ, NLP). Ειδικότερα, η ειρωνεία, ως φαινόμενο μη κυριολεκτικού λόγου, και ιδιαίτερα η ειρωνεία στον ελληνικό λόγο, αποτελεί ιδιαίτερη πρόκληση για τα υπολογιστικά συστήματα ανάλυσης κειμένου, καθώς απαιτούν από τα μοντέλα να κατανοούν όχι μόνο την κυριολεκτική σημασία των λέξεων, αλλά και τις πραγματολογικές προθέσεις του ομιλητή, τις κοινές εμπειρίες και το πολιτισμικό πλαίσιο σε αυτό που εκφέρει ο ομιλητής (Grice, 1975· Attardo, 2000).

Με την εκθετική ανάπτυξη των Μεγάλων Γλωσσικών Μοντέλων (ΜΓΜ, LLMs), ο τομέας της ΕΦΓ γνωρίζει σημαντική πρόοδο. Μοντέλα όπως το BERT και οι συνεχείς εξελίξεις τους έχουν αναδείξει την ικανότητα των υπολογιστών να αντιμετωπίζουν σύνθετα γλωσσολογικά φαινόμενα. Μια ουσιαστική ερώτηση όμως παραμένει αναπάντητη: σε τι βαθμό τα σύγχρονα ΜΓΜ είναι αποτελεσματικά σε γλώσσες χαμηλών πόρων και σε φαινόμενα που απαιτούν βαθιά πραγματολογική κατανόηση, όπως είναι η ειρωνεία στην ελληνική γλώσσα;

Η παρούσα εργασία στοχεύει να απαντήσει σε αυτό το ερώτημα μέσω μιας συγκριτικής εμπειρικής μελέτης. Αξιολογούνται τρεις προσεγγίσεις σε σύγκριση: ένας αλγόριθμος της Python, που βασίζεται σε κανόνες (rule-based) και εκφράζει την παραδοσιακή συμβολική λογική, και δύο σύγχρονα ΜΓΜ, το Claude και το DeepSeek, που αντιπροσωπεύουν την

τρέχουσα γενιά νευρωνικών μοντέλων. Η αξιολόγηση γίνεται σε σύγκριση με την ανθρώπινη επισήμανση σε 500 ελληνόγλωσσες διαδικτυακές κριτικές καταναλωτών. Αναμένεται να αποδειχθεί αν ο αλγόριθμος που βασίζεται σε κανόνες ή αν τα ΜΓΜ (Μεγάλα Γλωσσικά Μοντέλα) υπερτερούν στο ζήτημα αυτό, ενώ ταυτόχρονα αναδεικνύονται θεμελιώδεις διαφορές στον τρόπο με τον οποίο κάθε σύστημα αντιμετωπίζει την ειρωνική χρήση της γλώσσας, με ιδιαίτερη έμφαση στις ιδιαιτερότητες της ελληνικής γλώσσας ως γλώσσας χαμηλών πόρων. Το σώμα κειμένων της έρευνας αποτελείται από 500 διαδικτυακές κριτικές καταναλωτών αντλημένες από τις πλατφόρμες skroutz.gr και bestprice.gr, με επισημείωση ανθρώπινου αξιολογητή (πραγματική κατάσταση, ground truth). Η εστίαση στη γλώσσα κριτικών χρηστών αποτελεί ιδιαίτερα κατάλληλο πλαίσιο ερευνητικού ενδιαφέροντος, καθώς περιέχει φυσικό, μη επιμελημένο γραπτό λόγο με υψηλή συχνότητα σε στοιχεία ειρωνείας, σαρκασμού, χιούμορ και μεταφοράς (Maynard & Greenwood, 2014).

Ταυτόχρονα, οι διαδικτυακές κριτικές καταναλωτών συνιστούν ένα ιδιαίτερα απαιτητικό γλωσσικό περιβάλλον για την επεξεργασία φυσικής γλώσσας. Συνήθως είναι σύντομες και συχνά ανομοιογενείς ως προς τις γλωσσικές τους επιλογές: περιέχουν ανορθογραφίες, ελλείψεις τόνων, ή αντίθετα υπερβολική χρήση σημείων στίξης, όπως σειρές θαυμαστικών ή αποσιωπητικών, στοιχεία που δυσκολεύουν την αυτόματη επεξεργασία αλλά παράλληλα λειτουργούν συχνά ως φορείς ειρωνικής πρόθεσης (Davidov et al., 2010). Επιπλέον, οι κριτικές αυτές λειτουργούν χωρίς λεκτικό περικείμενο: ο αναγνώστης δεν γνωρίζει το ιστορικό της αγοράς, την προηγούμενη εμπειρία του χρήστη με το προϊόν, ούτε τον τόνο φωνής που συνόδευε τη γραφή. Σε αντίθεση με τον προφορικό λόγο, όπου η ειρωνεία σηματοδοτείται από τον τονισμό ή τις κινήσεις του ομιλητή, ο διαδικτυακός διάλογος παρέχει λιγοστές ενδείξεις και ο εντοπισμός της ειρωνείας εξαρτάται αποκλειστικά από εσωτερικούς κειμενικούς δείκτες (Pang & Lee, 2008).

Η ελληνική γλώσσα παρουσιάζει ιδιαίτερη πολυπλοκότητα στην επεξεργασία της με υπολογιστικό τρόπο, καθώς, μεταξύ άλλων, η πλούσια μορφολογία της, η ευελιξία στη σειρά των λέξεων και η εκτεταμένη χρήση υποκοριστικών με ειρωνική χροιά αποτελούν σειρά από προκλήσεις για τα συστήματα ΕΦΓ (Prokopidis & Piperidis, 2020). Επιπλέον, η ελληνική ανήκει στις γλώσσες χαμηλών πόρων (low-resource languages), με την έννοια ότι τα διαθέσιμα εκπαιδευτικά δεδομένα και αξιολογημένα σώματα κειμένων είναι περιορισμένα σε σύγκριση με κυρίαρχες γλώσσες όπως η αγγλική (Pitenis et al., 2021). Η συγκεκριμένη εργασία δεν

εστιάζει απλώς στην τεχνική ανίχνευση φαινομένων, αλλά διερευνά και τις θεμελιώδεις διαφορές μεταξύ των αλγοριθμικών παραδειγμάτων επεξεργασίας φυσικής γλώσσας, με στόχο την εξαγωγή συμπερασμάτων για την ανάπτυξη καλύτερων εργαλείων επεξεργασίας φυσικής γλώσσας για την ελληνική.

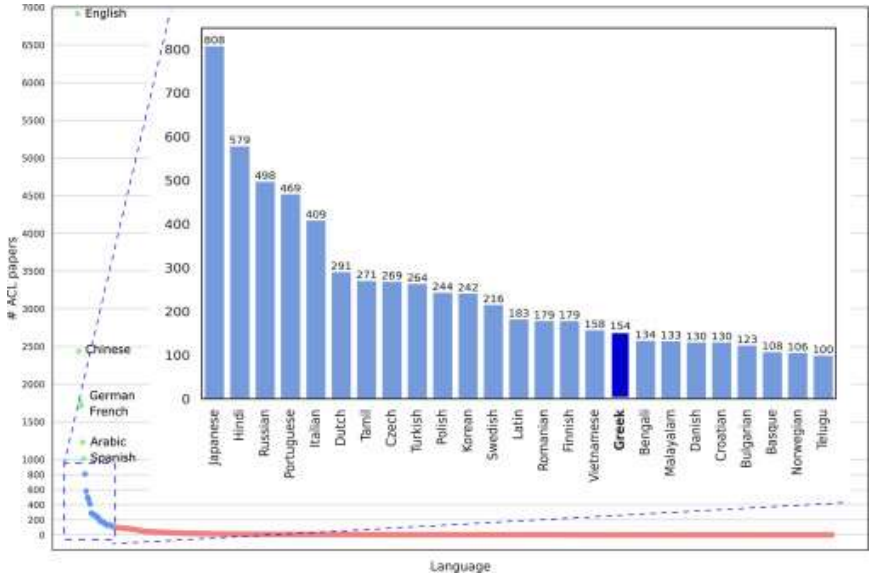
## 1.2. Ειρωνεία σε διαδικτυακές κριτικές

Η ειρωνεία αποτελεί κεντρικό στοιχείο της καθημερινής γλωσσικής έκφρασης. Το φαινόμενο λαμβάνει μεγαλύτερη έκταση ιδιαίτερα στο διαδικτυακό περιβάλλον, όπου οι χρήστες εκφράζουν απόψεις με υπονοούμενο και μερικές φορές ασαφή τρόπο. Στις διαδικτυακές κριτικές, η ειρωνεία λειτουργεί ως μηχανισμός έμμεσης αξιολόγησης. Ειδικότερα, ο συγγραφέας εκφράζει το αντίθετο από αυτό που εννοεί, αναμένοντας ο αναγνώστης να κατανοήσει το πραγματικό νόημα μέσα από το περικείμενο (Attardo, 2000). Αυτή η ασυμφωνία ανάμεσα στο κυριολεκτικό και το πραγματικό νόημα είναι ακριβώς αυτή που καθιστά την αυτόματη επεξεργασία της εξαιρετικά απαιτητική.

Η σημασία της αναγνώρισης αυτών των φαινομένων από τα γλωσσικά μοντέλα και τα λογισμικά ανάλυσης κειμένου είναι πολυδιάστατη. Καταρχάς, η μη ανίχνευση ειρωνείας οδηγεί σε εσφαλμένη ανάλυση συναισθήματος (sentimental analysis), με άμεσες επιπτώσεις στους σκοπούς διαδικασιών που σχετίζονται με την παρακολούθηση φήμης επιχειρήσεων, αξιολογήσεων έως και ακαδημαϊκές μελέτες κοινής γνώμης (Maynard & Greenwood, 2014). Μια κριτική που γράφει *«Εξαιρετική εξυπηρέτηση, περιμέναμε μόνο δύο ώρες»* θα κατηγοριοποιηθεί ως θετική από ένα μοντέλο που δεν αναγνωρίζει την ειρωνεία, ενώ στην πραγματικότητα εκφράζει έντονη δυσαρέσκεια.

Στο ελληνικό γλωσσικό περιβάλλον, το πρόβλημα οξύνεται από τη φύση της ελληνικής γλώσσας και από την απουσία εκτενών ερευνών στο συγκεκριμένο θέμα. Η πλούσια μορφολογία, η ευελιξία στη σειρά των λέξεων και η συχνή χρήση υποκοριστικών με ειρωνική χροιά (π.χ. *τέλειος, υπέροχος* σε σαρκαστικό πλαίσιο) δυσκολεύουν σημαντικά τα μοντέλα που έχουν εκπαιδευτεί κυρίως με αγγλικά δεδομένα (Prokopidis & Piperidis, 2020). Η ελληνική ανήκει στις λεγόμενες γλώσσες χαμηλών πόρων, δηλαδή στην κατηγορία γλωσσών που δεν διαθέτουν επαρκή ψηφιακά δεδομένα, σχολιασμένα σώματα κειμένων ή υπολογιστικά εργαλεία για την υποστήριξη της σύγχρονης ανάπτυξης ΕΦΓ, γεγονός που καθιστά ακόμα πιο επείγουσα την ανάπτυξη σχετικών πόρων και μεθοδολογιών (Pitenis et al., 2021). Όπως

φαίνεται στο Γράφημα 1 από το Bakagianni (2025), τα αγγλικά είναι η γλώσσα που έχει μελετηθεί περισσότερο όσον αφορά την ΕΦΓ στο περιβάλλον των ετήσιων συνεδρίων υπολογιστικής γλωσσολογίας, με 6.915 δημοσιεύσεις.




**Εικόνα 1 (Bakagianni, 2025): Αριθμός δημοσιεύσεων στην Ανθολογία ACL ανά γλώσσα (παρουσιάζεται κάθετα), με τις γλώσσες που αναφέρονται στον τίτλο ή στην περίληψη (οριζόντια)**

Τα κινέζικα, τα γερμανικά, τα γαλλικά, τα αραβικά και τα ισπανικά έχουν επίσης σημαντική κάλυψη, με αρκετές δημοσιεύσεις καθεμία. Τα ελληνικά ανήκουν στις γλώσσες με μέτρια υποστήριξη, με αριθμό δημοσιεύσεων που κυμαίνεται από 100 έως 1.000 ανά γλώσσα. Αναμένεται ότι μια παρόμοια εικόνα θα χαρακτηρίζει γενικά όλες τις δημοσιεύσεις στο πεδίο της ΕΦΓ.

Επιπλέον, σε ένα ευρύτερο κοινωνικό επίπεδο, η ικανότητα αυτόματης κατανόησης ειρωνείας συνδέεται άμεσα με την αξιοπιστία των ψηφιακών εργαλείων που χρησιμοποιούνται για την παρακολούθηση της κοινωνικής συζήτησης. Σε εποχή που τα ΜΓΜ εμπλέκονται ολοένα περισσότερο σε διαδικασίες λήψης αποφάσεων η αδυναμία κατανόησης της ειρωνείας μπορεί να έχει σοβαρές πρακτικές συνέπειες (Devlin et al., 2019).

## 1.3. Υπολογιστική ανίχνευση της ειρωνείας

Ως ειρωνεία ορίζεται η επικοινωνιακή πράξη κατά την οποία ο ομιλητής εκφέρει μια σκέψη που δεν ενστερνίζεται, προκειμένου να εκφράσει αρνητική στάση απέναντί της (Sperber & Wilson, 1981). Συχνά παρατηρείται σύγχυση μεταξύ ειρωνείας και σαρκασμού. Τα δεδομένα που αναλύονται περιλαμβάνουν κυρίως λεκτική ειρωνεία, παρόλα αυτά κατά πολλούς (Potamias et.all. 2019), η ειρωνεία αποτελεί «όρο-ομπρέλα» κάτω από την οποία βρίσκεται ο σαρκασμός. Κατά την έρευνα συμπεριλαμβάνουμε και τα δύο στον ανώτερο όρο, *ειρωνεία*.

Η ανίχνευση ειρωνείας στα σώματα κειμένων δεν αποτελεί αυτοσκοπό, αλλά κρίσιμο ενδιάμεσο βήμα σε μία σειρά από εφαρμογές επεξεργασίας φυσικής γλώσσας. Ο σημαντικότερος τομέας είναι η ανάλυση συναισθήματος, στην οποία στόχος είναι ο προσδιορισμός της συναισθηματικής πολικότητας ενός κειμένου ως θετικής, αρνητικής ή ουδέτερης. Τα σχήματα λόγου συχνά ανατρέπουν αυτή την πολικότητα και συνεπώς ένα σύστημα που δεν μπορεί να τα ανιχνεύσει θα παράγει συστηματικά λανθασμένα αποτελέσματα (Pang & Lee, 2008). Η ανάλυση συναισθήματος εφαρμόζεται σε πλήθος πλαισίων: αξιολόγηση κριτικών προϊόντων, παρακολούθηση κοινής γνώμης στα μέσα κοινωνικής δικτύωσης, ανάλυση πολιτικού λόγου και αξιολόγηση εμπειρίας πελατών (Liu, 2012).

Στο ηλεκτρονικό εμπόριο, πλατφόρμες όπως το Skroutz, το Amazon ή το TripAdvisor βασίζονται στην αυτόματη ανάλυση κριτικών για την παροχή συστάσεων και την κατάταξη προϊόντων ή υπηρεσιών. Εάν κριτικές με ειρωνικό περιεχόμενο ταξινομούνται εσφαλμένα, αυτό επηρεάζει άμεσα τόσο τους καταναλωτές όσο και τις επιχειρήσεις (Hu & Liu, 2004). Στην Ελλάδα, η αύξηση της διαδικτυακής αγοραστικής συμπεριφοράς, ιδίως μετά την πανδημία, έχει αναδείξει ακόμα περισσότερο την ανάγκη για αξιόπιστα εργαλεία ανάλυσης ελληνόγλωσσων κριτικών (Κουτσός & Αναγνωστόπουλος, 2022).

Στον τομέα της δημόσιας υγείας, η ανάλυση κοινωνικών μέσων χρησιμοποιείται για την παρακολούθηση στάσεων απέναντι σε θέματα όπως τα εμβόλια, οι θεραπείες ή οι δημόσιες πολιτικές υγείας. Ειρωνικές αναρτήσεις που εκφράζουν αντίθεση αλλά διαβάζονται ως υποστηρικτικές μπορούν να στρεβλώσουν σοβαρά την εικόνα που λαμβάνουν οι αρμόδιοι φορείς (Leaman et al., 2010). Αντίστοιχα, στη διαχείριση πολιτικών και κοινωνικών κρίσεων, η κατανόηση της ειρωνείας είναι απαραίτητη για τη σωστή αξιολόγηση του συναισθηματικού κλίματος του κοινού που θα οδηγήσει στον κατευνασμό και στη στοχευμένη διαχείριση της κοινωνίας σε περιόδους έκτακτης ανάγκης.

Στον ακαδημαϊκό χώρο, η ειρωνεία ως αντικείμενο μελέτης τέμνεται με πεδία όπως η υπολογιστική γλωσσολογία, η γνωσιακή επιστήμη και η κοινωνιογλωσσολογία, δίνοντας έτσι πολυεπίπεδη σημασία στην έρευνα (Gibbs, 2000). Ιδιαίτερη θέση στη μελέτη της ειρωνείας κατέχουν παραδοσιακά η Πραγματολογία και η Ανάλυση Λόγου. Η Πραγματολογία προσεγγίζει την ειρωνεία ως φαινόμενο που δεν μπορεί να ερμηνευθεί βάσει του κυριολεκτικού νοήματος των λέξεων, αλλά μόνο μέσα από την κατανόηση της επικοινωνιακής πρόθεσης του ομιλητή και του πλαισίου της εκφοράς, μια προσέγγιση που θεμελιώθηκε από τον Grice (1975) με τη θεωρία των συνομιλιακών μέγιστων και επεκτάθηκε από τους Sperber & Wilson (1981) στο πλαίσιο της Θεωρίας Συνάφειας. Η Ανάλυση Λόγου, από την πλευρά της, εξετάζει την ειρωνεία ως κοινωνική πρακτική, δηλαδή, ως τρόπο διαχείρισης ταυτοτήτων, σχέσεων εξουσίας και κοινωνικών νορμών μέσα σε συγκεκριμένα επικοινωνιακά πλαίσια (Fairclough, 1992· Culpeper et al., 2017). Και οι δύο διαστάσεις είναι σχετικές για την υπολογιστική ανίχνευση, καθώς ένα σύστημα που αγνοεί την πραγματολογική διάσταση δεν μπορεί να κατανοήσει την ειρωνεία, ενώ ένα που αγνοεί την κοινωνική της λειτουργία δεν μπορεί να εκτιμήσει γιατί εμφανίζεται και σε ποιο πλαίσιο (Attardo, 2000). Η ελληνική ακαδημαϊκή κοινότητα έχει αρχίσει να ασχολείται με το πεδίο, κυρίως με έρευνες για την αυτόματη επεξεργασία της ελληνικής γλώσσας και την ανάπτυξη σχετικών γλωσσικών πόρων (Markantonatou et al., 2019).

Οι προκλήσεις που αντιμετωπίζουν τα αυτόματα συστήματα κατά την ανάλυση κριτικών είναι πολυεπίπεδες και αφορούν τόσο γλωσσικές όσο και υπολογιστικές δυσκολίες. Η πρώτη και πιο θεμελιώδης πρόκληση είναι η έλλειψη εξωγλωσσικού περικειμένου. Για παράδειγμα, η ειρωνεία στον προφορικό λόγο πραγματώνεται μέσα από τον τόνο της φωνής, τη γλώσσα του σώματος και τις κοινές εμπειρίες των συνομιλητών, στοιχεία που απουσιάζουν εντελώς από το γραπτό κείμενο και ιδιαίτερα από το κείμενο μιας κριτικής (Attardo, 2000).

Ακόμα μια σημαντική πηγή σφάλματος, ιδιαίτερα σε κριτικές προϊόντων, αποτελεί η ασυνέπεια μεταξύ αριθμητικής βαθμολογίας και γλωσσικού περιεχομένου. Είναι συνηθισμένο φαινόμενο χρήστες να δίνουν ένα ή δύο αστέρια με ειρωνικά «επαινετικό» κείμενο, ή αντίστροφα να δίνουν πέντε αστέρια συνοδεύοντάς τα με σαρκαστικό σχόλιο. Αυτή η αντίφαση συνήθως οδηγεί σε εσφαλμένη εκπαίδευση και ταξινόμηση (Davidov et al., 2010). Το γλωσσικό μοντέλο λαμβάνει υπόψη μόνο το κείμενο κριτικής και όχι την τελική βαθμολογία.

Στο πεδίο της ελληνικής γλώσσας, ιδιαίτερη πρόκληση παρουσιάζει η πλούσια μορφολογία και η συντακτική ευελιξία. Η ελληνική επιτρέπει πολλαπλές διατάξεις λέξεων στην πρόταση χωρίς αλλαγή της βασικής σημασίας, κάτι που δυσκολεύει τα μοντέλα να εντοπίσουν ειρωνικές δομές (Prokopidis & Piperidis, 2020). Επιπλέον, η χρήση υποκοριστικών με ειρωνική λειτουργία (π.χ. *καλούτσικο, ωραιάκι*) αποτελεί γνώρισμα της ελληνικής καθημερινής γλώσσας που σπάνια αναπαρίσταται στα εκπαιδευτικά σύνολα δεδομένων.

Επιπρόσθετα, τα ορθογραφικά και τυπογραφικά λάθη των χρηστών (π.χ. *τελεοιο, εξεριτικο*) παρεμποδίζουν την κατάτμηση σε δείγματα (tokenization) και αναγνώριση λέξεων που οι περισσότεροι αλγόριθμοι δεν έχουν εκπαιδευτεί σε τέτοιο βαθμό ώστε να είναι σε θέση να τα αντιμετωπίσουν (Chalamandaris et al., 2006).

Η υπερβολή και η λιτότητα ως μορφές ειρωνικής έκφρασης προσθέτουν επιπλέον πολυπλοκότητα, όπως επίσης το ίδιο συμβαίνει και με την χρήση των σημείων στίξεως όπως το θαυμαστικό. Προτάσεις όπως *Το καλύτερο σουβλάκι στον πλανήτη!* ή *Όχι και τόσο άσχημο…* απαιτούν από το μοντέλο να αξιολογήσει τον βαθμό υπερβολής σε σχέση με το κανονικό εύρος αξιολογήσεων για παρόμοια προϊόντα, κάτι που απαιτεί βαθύτερη κατανόηση πραγματολογικών κανόνων (Veale & Hao, 2010).

Τέλος, η πολιτισμική διακύμανση στην έκφραση ειρωνείας σημαίνει ότι αυτό που θεωρείται σαρκαστικό σε ένα πολιτισμικό πλαίσιο μπορεί να εκλαμβάνεται κυριολεκτικά σε ένα άλλο (Gibbs, 2000). Αυτό δυσχεραίνει τον εντοπισμό της ειρωνείας σε ένα διαφορετικό πλαίσιο αλλά και γενικότερα τη γενίκευση των μοντέλων.

## 1.4. Ερευνητικά ερωτήματα

Σκοπός της παρούσας εργασίας είναι η συγκριτική αξιολόγηση τριών διαφορετικών υπολογιστικών προσεγγίσεων ως προς την ικανότητά τους να ανιχνεύουν ειρωνεία σε ελληνόγλωσσες διαδικτυακές κριτικές καταναλωτών: έναν αλγόριθμο βασισμένο σε κανόνες υλοποιημένο σε Python και δύο σύγχρονα Μεγάλα Γλωσσικά Μοντέλα, το Claude (Anthropic) και το DeepSeek (DeepSeek AI). Η συγκριτική αυτή ανάλυση στοχεύει να αναδείξει τις δυνατότητες και τα όρια κάθε παραδείγματος επεξεργασίας, να απαντήσει σε ερευνητικά ερωτήματα για την ελληνική γλώσσα ως γλώσσα λίγων γλωσσικών πόρων, να εξετάσει την ικανότητα διαγλωσσικής μεταφοράς μάθησης (cross-lingual transfer learning) των ΜΓΜ και

να συνεισφέρει στον πιθανό ευρύτερο προβληματισμό της αναξιοπιστίας των διαδικτυακών εργαλείων επεξεργασίας φυσικής γλώσσας για την ελληνική.

Ειδικότερα, τίθενται τα εξής ερευνητικά ερωτήματα:

ΕΡ1: Πόσο ακριβής είναι ο βασισμένος σε κανόνες αλγόριθμος στην ανίχνευση ειρωνείας σε ελληνόγλωσσες διαδικτυακές κριτικές καταναλωτών;

ΕΡ2: Πώς αποδίδουν τα Μεγάλα Γλωσσικά Μοντέλα που επιλέχθηκαν στην ίδια εργασία ανίχνευσης, σε σύγκριση με τον αλγόριθμο αλλά και μεταξύ τους;

ΕΡ3: Ποια γλωσσικά χαρακτηριστικά της ελληνικής γλώσσας (ιδιωματισμοί, υποκοριστικά με ειρωνική χροιά, συντακτική ευελιξία) δυσκολεύουν την αυτόματη ανίχνευση ειρωνείας;

ΕΡ4: Υπάρχει συστηματική διαφορά στον τρόπο με τον οποίο κάθε σύστημα αντιμετωπίζει την ειρωνεία, και ποια είναι τα τυπικά γλωσσικά δομικά σχήματα λαθών κάθε προσέγγισης;

Η παρούσα εργασία αναμένεται να αναδείξει τι από τα δύο υπερτερεί: τα Μεγάλα Γλωσσικά Μοντέλα ή η προσέγγιση που βασίζεται σε κανόνες στην ανίχνευση ειρωνείας σε ελληνόγλωσσα κείμενα; Η έρευνα ακόμα θα αναδείξει την ικανότητά τους να χειρίζονται τη μη κυριολεκτική γλώσσα ακόμα και σε γλώσσα χαμηλών πόρων. Παράλληλα, αναμένεται να διαπιστωθεί ότι τα δύο ΜΓΜ θα εμφανίζουν διαφορετικές στρατηγικές ανίχνευσης της ειρωνείας σε σύγκριση μεταξύ τους.

# 2. Θεωρητικό πλαίσιο

Το δεύτερο κεφάλαιο παρέχει το θεωρητικό και βιβλιογραφικό υπόβαθρο πάνω στο οποίο στηρίζεται η ερευνητική σχεδίαση της εργασίας. Αρχικά, περιγράφεται η ερευνητική πορεία της αυτόματης ανίχνευσης ειρωνείας στην υπολογιστική γλωσσολογία, από τις πρώτες προσεγγίσεις με αλγορίθμους βασισμένους σε κανόνες έως τα σύγχρονα νευρωνικά μοντέλα (2.1). Στη συνέχεια παρουσιάζονται οι κυριότερες πρακτικές και επιστημονικές στρατηγικές αντιμετώπισης των προκλήσεων που θέτει η ανίχνευση μη κυριολεκτικής γλώσσας, με έμφαση στις προσεγγίσεις που έχουν αναπτυχθεί για γλώσσες χαμηλών πόρων (2.2). Το κεφάλαιο ολοκληρώνεται με εστιασμένη ανασκόπηση της πρόσφατης βιβλιογραφίας σχετικά με τη χρήση ΜΓΜ για τον εντοπισμό ειρωνείας, αναδεικνύοντας τα εμπειρικά ευρήματα και τα ανοιχτά ερωτήματα που η παρούσα εργασία καλείται να διερευνήσει (2.3).

## 2.1. Γλωσσολογική ανάλυση της ειρωνείας

### 2.1.1 Η ειρωνεία στο πλαίσιο της θεωρίας του Grice: Παραβίαση του αξιώματος της ποιότητας

Το έργο του H. P. Grice (1975/1989), αποτελεί μία από τις πιο συστηματικές πραγματολογικές αναλύσεις της ειρωνείας. Η θεωρία του για τις συνομιλιακές συνεπαγωγές αποτελεί τομή για την σύγχρονη έρευνα. Σύμφωνα με τον Grice, η επικοινωνία διέπεται από μία υπερκείμενη αρχή, την Αρχή Συνεργασίας (Cooperative Principle), η οποία αναλύεται σε τέσσερις επιμέρους Αξιώματα: Ποιότητα, Ποσότητα, Συνάφεια και Τρόπο (Γούτσος, 2020). Η ειρωνεία, κατά τον Grice, είναι μια σημαντική παραβίαση της πρώτης, της Ποιότητας. Αναλυτικότερα όπως διατυπώνει ρητά: «Μη λες κάτι που πιστεύεις ότι είναι ψευδές» (Culpeper et al., 2017, σ. 325).

Το χαρακτηριστικό παράδειγμα που χρησιμοποιεί ο Grice είναι η φράση «Ο Χ είναι εξαιρετικός φίλος» («X is a fine friend»), η οποία εκφέρεται σε περίσταση όπου ο Χ έχει μόλις προδώσει ένα μυστικό. Σε αυτή την περίπτωση, ο ακροατής αντιλαμβάνεται ότι αυτό που ειπώθηκε αντιβαίνει στην πραγματικότητα και συμπεραίνει πως ο ομιλητής ηθελημένα παραβιάζει τη Ποιότητα ώστε να μεταδώσει το αντίθετο νόημα, δηλαδή ότι ο Χ δεν είναι καλός φίλος (Culpeper et al., 2017, σ. 325). Η λογική αυτής της ανάλυσης στηρίζεται στην έννοια της αντίθεσης: το ειρωνικό νόημα είναι η άρνηση του κυριολεκτικού.

Ωστόσο, ο ίδιος ο Grice αναγνώρισε ότι η ανάλυση αυτή ήταν ελλιπής. Σε μεταγενέστερο κείμενό του επεσήμανε ότι η ειρωνεία συνδέεται άρρηκτα με την έκφραση ενός αισθήματος ή μιας στάσης (συνήθως έντονης), η οποία αποτυπώνεται μέσα από έναν «ειρωνικό τόνο» (ironical tone), που σχετίζεται κατά κύριο λόγο με προσωδιακά χαρακτηριστικά (Culpeper et al., 2017, σ. 326). Επίσης, υποστήριξε ότι η ειρωνεία εμπεριέχει στοιχείο προσποίησης, καθώς ο ομιλητής «προσποιείται» ότι λέει κάτι που στην πραγματικότητα δεν πιστεύει.

Η προσέγγιση αυτή, αν και θεμελιώδης για τη μελέτη της ειρωνείας, έχει δεχθεί κριτική για αρκετές αδυναμίες. Αρχικά, η έμφαση στη σχέση αντίθεσης δεν εξηγεί όλα τα είδη ειρωνείας, όπως η υπερβολή και η λιτότητα, που δεν στηρίζονται απαραίτητα σε λογική αντίφαση (Culpeper et al., 2017, σ. 374). Επίσης, ο μηχανισμός παραγωγής συνεπαγωγής περιγράφει τι συμβαίνει χωρίς να εξηγεί επαρκώς πώς ο ακροατής φτάνει στο ειρωνικό νόημα και γιατί ορισμένες ερμηνείες είναι πιο πιθανές από άλλες.

### 2.1.2 Η ερμηνεία της ηχώς (echoic account) στο πλαίσιο της Θεωρίας της Συνάφειας

Η πιο επιδραστική εναλλακτική πρόταση για την ανάλυση της ειρωνείας προέρχεται από τη Θεωρία Συνάφειας (Relevance Theory) των Sperber και Wilson (1981, 1995), η οποία εισήγαγε την ερμηνεία της ηχώς ή μιμητική ερμηνεία (echoic account). Η θεωρία αυτή αντιμετωπίζει την ειρωνεία ως ένα φαινόμενο που εμπλέκει την απόδοση σκέψεων (attributed thoughts) σε άλλα πρόσωπα, ομάδες και την ταυτόχρονη έκφραση αποστασιοποιημένης στάσης απέναντί τους.

Με βάση το παραπάνω παράδειγμα στο 2.1.1 «Τι εξαιρετικός φίλος που είσαι!» μετά από μια απογοήτευση ο ομιλητής δεν ισχυρίζεται ούτε αρνείται ότι ο άλλος είναι καλός φίλος. Αυτό που κάνει είναι να ανακαλεί όπως η ηχώ την κοινώς αποδεκτή προσδοκία ότι οι φίλοι συμπεριφέρονται με αφοσίωση και να εκφράζει παράλληλα την απογοήτευσή του για το γεγονός ότι αυτή η προσδοκία δεν ικανοποιήθηκε (Wilson & Sperber, 2012, βλ. και Culpeper et al., 2017, σ. 374). Η ειρωνεία, με άλλα λόγια, δεν είναι αυτό που λέγεται, αλλά αυτό που ανακαλείται και απορρίπτεται ταυτόχρονα.

Ένα κεντρικό στοιχείο της θεωρίας είναι ο ρόλος των κοινωνικών νορμών. Οι Wilson και Sperber (2012, βλ. Culpeper et al., 2017, σ. 375) επισημαίνουν ότι οι νόρμες αποτελούν ανεξάντλητη πηγή ειρωνικής ανάκλησης, καθώς στην καθημερινή ζωή παραβιάζονται συνεχώς. Αυτό εξηγεί γιατί εκφράσεις όπως «Τι ωραίος καιρός!» σε συνθήκες κακοκαιρίας ή «Μπράβο σου!» μετά από αποτυχία γίνονται εύκολα κατανοητές ως ειρωνικές χωρίς να χρειάζεται περαιτέρω επεξήγηση: ο ακροατής αναγνωρίζει αυτόματα ότι η νόρμα παραβιάστηκε και ότι ο ομιλητής το σχολιάζει έμμεσα με ειρωνεία ή σαρκασμό.

Η στάση που εκφράζεται μέσω της ειρωνείας κυμαίνεται σε ευρύ φάσμα: από ήπια αυταρέσκεια ή χιούμορ ως απογοήτευση, περιφρόνηση ή οργή (Wilson & Sperber, 2012, σ. 130, Culpeper et al., 2017, σ. 374). Αυτή η ευελιξία καθιστά τη θεωρία ικανή να αποδώσει τόσο τις ήπιες, παιχνιδιάρικες μορφές ειρωνείας όσο και τον σαρκασμό που στοχεύει συγκεκριμένο πρόσωπο ή κατάσταση (Culpeper et al., 2017, σ. 327).

Σε επίπεδο γνωστικής επεξεργασίας, η Wilson (2009, βλ. Huang, 2017, σ. 100) υποστηρίζει ότι η κατανόηση ειρωνείας απαιτεί υψηλότερης τάξης κατανόηση των νοητικών λειτουργιών (mindreading) σε σύγκριση με κυριολεκτικές ή μεταφορικές εκφράσεις, δηλαδή, προϋποθέτει την ικανότητα να «μαντεύει» κανείς την πρόθεση του ομιλητή πίσω από αυτό που λέγεται. Αυτό επιβεβαιώνεται εμπειρικά από μελέτες που καταγράφουν δυσκολίες κατανόησης ειρωνείας σε περιπτώσεις εγκεφαλικής βλάβης (Huang, 2017, σ. 509), και έχει άμεση συνέπεια για την αυτόματη ανίχνευσή της: αν η ειρωνεία απαιτεί ακόμα και για τον άνθρωπο υψηλότερου επιπέδου γνωστική επεξεργασία, είναι αναμενόμενο να αποτελεί ιδιαίτερη πρόκληση για κάθε υπολογιστικό σύστημα που επεξεργάζεται γλώσσα σε επίπεδο επιφανειακών μοτίβων.

### 2.1.3 Η υπόθεση της βαθμιαίας προεξοχής (Graded Salience Hypothesis) και γνωστικές προσεγγίσεις

Παράλληλα με τις πραγματολογικές θεωρίες, η ψυχογλωσσολογική έρευνα επικεντρώθηκε στον τρόπο επεξεργασίας της ειρωνείας. Το κλασικό μοντέλο που θέτει ως προτεραιότητα την κυριολεκτική σημασία της γλώσσας (literal-first model) υποστηρίζει ότι η κατανόηση μη κυριολεκτικής γλώσσας γίνεται σε δύο βήματα: αρχικά ανακτάται το κυριολεκτικό νόημα και στη συνέχεια, αν κριθεί ακατάλληλο για το πλαίσιο, αναζητείται το πρόσθετο νόημα. Ο Gibbs

(1994, βλ. Huang, 2017, σ. 27) αμφισβήτησε αυτό το μοντέλο, υποστηρίζοντας ότι η πραγματολογική επεξεργασία είναι παράλληλη και όχι σταδιακή.

Η Rachel Giora (2003, 2004, βλ. Huang, 2017, σ. 27) πρόσθεσε μία σημαντική διάσταση σε αυτή την ερμηνεία με την *υπόθεση της βαθμιαίας προεξοχής*. Σύμφωνα με αυτή, αυτό που καθορίζει την τάξη επεξεργασίας δεν είναι η κυριολεκτικότητα αλλά η προεξοχή (salience): το πιο εμφανές νόημα, είτε κυριολεκτικό είτε μη, ανακτάται πρώτο. Για ειρωνικές εκφράσεις το ειρωνικό νόημα έχει υψηλή προεξοχή και ανακτάται άμεσα, ενώ για λεπτές και «κρυμμένες» ειρωνείες απαιτείται μεγαλύτερος χρόνος επεξεργασίας (Huang, 2017, σ. 27–28).

Συνοψίζοντας, οι γλωσσολογικές θεωρίες της ειρωνείας εξελίχθηκαν από μία απλοϊκή λογική αντίθεσης (Grice) σε λεπτομερείς αναλύσεις που ενσωματώνουν τη απόδοση νοητικών λειτουργιών στον συνομιλητή (Sperber & Wilson) ή τη γνωστική προεξοχή (Giora), αλλά και σε συνάρτηση και με τον ρόλο που παίζουν οι κοινωνικές νόρμες. Κάθε προσέγγιση φωτίζει διαφορετικές πτυχές του φαινομένου: η ανάλυση του Grice αναδεικνύει τον χαρακτήρα της ειρωνείας που βασίζεται σε συλλογισμούς, η ερμηνεία της ηχώς αναφέρεται στον κοινωνικό της χαρακτήρα σε σχέση με τη στάση του ομιλητή, ενώ οι γνωστικές προσεγγίσεις υπογραμμίζουν την πολυπλοκότητα της επεξεργασίας της. Σήμερα, τα πεδία έρευνας που παρουσιάζουν τη μεγαλύτερη δυναμική αφορούν τη διαπολιτισμική μελέτη της ειρωνείας, τον ρόλο της πολυτροπικότητας (π.χ. τόνος φωνής, μιμική) και την αλληλεπίδρασή της με τις έννοιες ευγένειας και αγένειας (Culpeper et al., 2017, σ. 380).

## 2.2 Τι είναι τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ);

Τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ, Large Language Models, LLMs), αποτελούν μια κατηγορία νευρωνικών δικτύων που εκπαιδεύονται σε τεράστια σώματα κειμένων με στόχο την κατανόηση και παραγωγή φυσικής γλώσσας. Η αρχιτεκτονική τους βασίζεται στον μηχανισμό προσοχής (attention), που εισήχθη από τους Vaswani et al. (2017) με το μοντέλο Transformer και επιτρέπει στο σύστημα να «εστιάζει» σε σχετικά τμήματα του κειμένου κατά την επεξεργασία, αντί να το διαβάζει γραμμικά. Αυτό αποτέλεσε τομή στο πεδίο και εισήγαγε μια θεμελιώδη αλλαγή σε σχέση με τις προηγούμενες αρχιτεκτονικές, καθώς επέτρεψε για πρώτη φορά την αποτελεσματική μοντελοποίηση μακροπρόθεσμων γλωσσικών εξαρτήσεων.

Η μεγάλη τομή ήρθε με την οικογένεια μοντέλων BERT (Devlin et al., 2019) και GPT (Radford et al., 2019), τα οποία απέδειξαν ότι η προεκπαίδευση σε μεγάλα σώματα κειμένων, χωρίς εποπτεία και χωρίς εξειδίκευση σε συγκεκριμένη εργασία, ακολουθούμενη από προσαρμογή (fine-tuning) σε μικρότερα εξειδικευμένα δεδομένα, οδηγεί σε ανώτερα αποτελέσματα σε ένα ευρύ φάσμα γλωσσικών εργασιών. Τα σύγχρονα ΜΓΜ επεκτείνουν αυτή τη λογική σε πολύ μεγαλύτερη κλίμακα, καθώς πλέον λαμβάνουν υπόψη δισεκατομμύρια παραμέτρους και τρισεκατομμύρια λέξεις εκπαίδευσης, με αποτέλεσμα να αναπτύσσουν ικανότητες που υπερβαίνουν τις εργασίες για τις οποίες εκπαιδεύτηκαν ρητά (Brown et al., 2020).

Μια κεντρική ιδιότητα των σύγχρονων ΜΓΜ είναι η ικανότητά τους να εκτελούν εργασίες χωρίς εξειδικευμένη εκπαίδευση, απλώς μέσω κατάλληλης διατύπωσης οδηγιών, της λεγόμενης προτροπής (prompt, Brown et al., 2020). Αυτή η ιδιότητα τα καθιστά ιδιαίτερα ελκυστικά για εφαρμογές σε γλώσσες χαμηλών πόρων, όπως η ελληνική, όπου η διαθεσιμότητα εξειδικευμένων εκπαιδευτικών δεδομένων είναι περιορισμένη (Conneau et al., 2020). Ωστόσο, η απόδοσή τους σε τέτοιες γλώσσες εξαρτάται κρίσιμα από τον βαθμό εκπροσώπησής τους στα δεδομένα προεκπαίδευσης, γεγονός που δημιουργεί δομικές ανισότητες μεταξύ γλωσσών (Pitenis et al., 2021).

Στην παρούσα εργασία αξιολογούνται δύο ΜΓΜ με διαφορετική καταγωγή και στρατηγική εκπαίδευσης: το Claude (Anthropic), που έχει αναπτυχθεί με έμφαση στο Constitutional AI και την ανθρώπινη ανατροφοδότηση RLHF (Bai et al., 2022) και το DeepSeek (DeepSeek AI), που προέρχεται από κινεζική ερευνητική ομάδα με διαφορετική προσέγγιση εκπαίδευσης και δεδομένα (DeepSeek-AI, 2024). Η επιλογή δύο τεχνολογικά και πολιτισμικά διαφορετικών μοντέλων δεν είναι τυχαία: αποσκοπεί στο να αναδείξει αν και κατά πόσο οι διαφορές στη στρατηγική εκπαίδευσης επηρεάζουν τον τρόπο με τον οποίο κάθε μοντέλο προσεγγίζει ένα πολιτισμικά φορτισμένο φαινόμενο όπως η ειρωνεία στην ελληνική γλώσσα.

## 2.3 Ειρωνεία στην υπολογιστική γλωσσολογία

Η ερευνητική δραστηριότητα σχετικά με την αυτόματη ανίχνευση ειρωνείας γνωρίζει σημαντική ανάπτυξη την τελευταία δεκαετία. Οι πρωτοπόρες συστηματικές απόπειρες στην ανίχνευση ειρωνείας ακολουθούσαν αλγορίθμους βασισμένους σε κανόνες: συνδυασμούς λεξικών δεικτών (θετικών λέξεων με αρνητικό συναίσθημα) και στατιστικών συσχετισμών (Riloff et al., 2013). Οι σημαντικές εργασίες των Davidov et al. (2010) και Maynard & Greenwood (2014) κατέδειξαν ότι η ειρωνεία αποτελεί συστηματική πηγή σφαλμάτων στην ανάλυση συναισθήματος, καθώς τα συστήματα εκλαμβάνουν εξ ορισμού λανθασμένα την ειρωνεία ως θετικό σημείο.

Η άνοδος των νευρωνικών αρχιτεκτονικών και προεκπαιδευμένων μοντέλων (Devlin et al., 2019) άνοιξε νέους ορίζοντες για την ανίχνευση μη κυριολεκτικής γλώσσας και ειρωνείας. Οι Brown et al. (2020) υποστήριξαν ότι τα ΜΓΜ εμφανίζουν ικανοποιητικά αποτελέσματα σε εργασίες κατανόησης γλώσσας χωρίς εξειδίκευση few-shot ή zero-shot: στην *προσέγγιση με λίγα παραδείγματα* (few-shot), το μοντέλο καθοδηγείται μέσα από έναν μικρό αριθμό παραδειγμάτων που ενσωματώνονται απευθείας στην προτροπή (συνήθως δύο έως πέντε παραδείγματα) τα οποία υποδεικνύουν τον τύπο της απάντησης που αναμένεται, χωρίς ωστόσο να τροποποιούνται οι παράμετροι του μοντέλου. Η διαφορά είναι ότι στην *διαδικασία μηδενικής μάθησης* (τεχνική zero-shot), το μοντέλο «μαντεύει» βάσει γενικής κατανόησης, ενώ στην προσέγγιση με λίγα παραδείγματα «βλέπει» την θεωρία που του έχει δοθεί πριν απαντήσει (Brown et al., 2020).

Οι Oprea και Bâra (2025) υπογράμμισαν ότι ο σχεδιασμός των προτροπών αποτελεί κρίσιμο παράγοντα για την απόδοση των μοντέλων, καταλήγοντας ότι τα ΜΓΜ τείνουν να υπακούν στις κατευθύνσεις της ερώτησης. (Το συμπέρασμα αυτό συνάδει άμεσα με το εύρημα της παρούσας εργασίας σχετικά με τη μεροληψία των ΜΓΜ προς την κατεύθυνση των προτροπών, όπως θα διαπιστώσουμε αργότερα).

Η ερευνητική βιβλιογραφία αναδεικνύει ότι η σύγκριση διαφορετικών υπολογιστικών προσεγγίσεων για την ανίχνευση ειρωνείας αποτελεί ενεργό ερευνητικό πεδίο. Ο Ortega Bueno (2025) εκπόνησε εκτεταμένη διδακτορική διατριβή σχετικά με νέες μεθόδους για τον εντοπισμό ειρωνείας σε κοινωνικά μέσα, επισημαίνοντας ότι η γλωσσική και πολιτισμική

ιδιαιτερότητα της ειρωνείας απαιτεί εξειδικευμένα σώματα κειμένων ανά γλώσσα. Η παρούσα εργασία ευθυγραμμίζεται με αυτή την ερευνητική άποψη, προσθέτοντας την ελληνική γλώσσα ως πεδίο ερευνητικού ενδιαφέροντος και καταγράφοντας τις προκλήσεις για τα υφιστάμενα ΜΓΜ.

Κεντρικό έναυσμα της παρούσας εργασίας αποτελεί η διαπίστωση ότι η ανίχνευση ειρωνείας δεν είναι απλώς τεχνικό πρόβλημα, αλλά εννοιακό και γλωσσικό. Αναλυτικότερα, το ζήτημα απαιτεί από τα μοντέλα να κατανοούν όχι μόνο το κείμενο αλλά και το περικείμενο, το πραγματολογικό πλαίσιο, τις πολιτισμικές συμβάσεις και την κοινή γνώση των ελληνοφώνων στη συγκεκριμένη περίπτωση (Gibbs, 2000).

## 2.4 Προσεγγίσεις βελτίωσης συστημάτων εντοπισμού της ειρωνείας

Η αντιμετώπιση των παραπάνω προκλήσεων απαιτεί συνδυασμό πρακτικών παρεμβάσεων και βαθύτερων επιστημονικών καινοτομιών. Σε πρακτικό επίπεδο, η πιο άμεσα εφαρμόσιμη βελτίωση είναι η ενσωμάτωση μεταδεδομένων στη διαδικασία ανάλυσης. Η αριθμητική βαθμολογία (αστέρια) που συνοδεύει μια κριτική μπορεί να αποτελέσει πολύτιμο σηματοδότη. Ειδικότερα, η ασυμφωνία μεταξύ βαθμολογίας και γλωσσικού τόνου αποτελεί ισχυρή ένδειξη ειρωνείας (Davidov et al., 2010). Αντίστοιχα, η ανάλυση εικονιδίων emojis και σημείων στίξης (π.χ. θαυμαστικά, εισαγωγικά) παρέχει μια επιπλέον πλαισίωση που αγνοούν τα αμιγώς κειμενικά μοντέλα (Poria et al., 2016).

Σε τεχνολογικό-ερευνητικό επίπεδο, η χρήση προ-εκπαιδευμένων πολύγλωσσων μοντέλων όπως το mBERT (multilingual BERT) ή το XLM-RoBERTa έχει αποδειχθεί ιδιαίτερα υποσχόμενη για γλώσσες με περιορισμένους πόρους όπως η ελληνική, καθώς επιτρέπει τη μεταφορά γνώσης από γλώσσες με πλούσια δεδομένα (Conneau et al., 2020).

Σε πιο προχωρημένο επιστημονικό επίπεδο, η πολυτροπική ανάλυση (multimodal analysis) αξιοποιεί συνδυασμό κειμένου, εικόνας και ηχητικού σήματος για ολιστική κατανόηση του νοήματος (Poria et al., 2016). Επιπλέον, η χρήση μηχανισμών προσοχής (attention mechanisms) και γράφων γνώσης επιτρέπει στα μοντέλα να αντλούν πολιτισμικό και εγκυκλοπαιδικό υπόβαθρο, ενισχύοντας την ικανότητά τους να κατανοούν συγκεκριμένες ειρωνικές αναφορές που προϋποθέτουν κοινή γνώση (Veale & Hao, 2010).

Η έρευνα για την αυτόματη ανίχνευση ειρωνείας μέσω μεγάλων γλωσσικών μοντέλων βρίσκεται σε εξέλιξη, με τα τελευταία χρόνια να σηματοδοτούν σαφή στροφή από τις παραδοσιακές μεθόδους που βασίζονται σε κανόνες και επιτήρηση (rule-based και supervised) σε μια προσέγγιση με Μηδενική Εκμάθηση (zero-shot) και Εκμάθηση με Λίγα Παραδείγματα (few-shot) προσεγγίσεις βασισμένες σε προεκπαιδευμένα μοντέλα (Brown et al., 2020; Wei et al., 2022).

Η θεμελιώδης πρόκληση για κάθε σύστημα είναι ότι η ειρωνεία δεν αποτελεί απλή λεξιλογική κατηγορία αλλά πραγματολογικό φαινόμενο που απαιτεί κατανόηση της πρόθεσης του ομιλητή, του περικειμένου και της κοινής γνώσης (Grice, 1975; Sperber & Wilson, 1981). Αυτή η ιδιότητα καθιστά θεωρητικά τα ΜΓΜ κατάλληλα για τον εντοπισμό της ειρωνείας, καθώς εκπαιδεύονται σε τεράστια σώματα κειμένων που ενσωματώνουν πραγματολογικές και πολιτισμικές συμβάσεις (Devlin et al., 2019).

Εμπειρικά, ωστόσο, τα αποτελέσματα παραμένουν ανομοιογενή. Οι Brown et al. (2020) απέδειξαν ότι τα ΜΓΜ μεγάλης κλίμακας εμφανίζουν αξιοσημείωτες ικανότητες σε εργασίες κατανόησης γλώσσας με τεχνικές με μηδενικά παραδείγματα (zero-shot), συμπεριλαμβανομένης της ανίχνευσης ειρωνικής έκφρασης. Ένα επαναλαμβανόμενο εύρημα στη βιβλιογραφία αφορά την ασυμμετρία μεταξύ Ακρίβειας (Precision) και Ανάκλησης (Recall) στα ΜΓΜ. Τα μοντέλα που εκπαιδεύονται με έμφαση στην ασφάλεια και την αποφυγή λαθών (όπως το Claude μέσω Constitutional AI και RLHF) τείνουν προς υψηλότερη ανάκληση, δηλαδή εντοπίζουν περισσότερες ειρωνικές περιπτώσεις αλλά με κόστος αυξημένων ψευδώς θετικών, ενώ μοντέλα με διαφορετική στρατηγική εκπαίδευσης εμφανίζουν το αντίθετο προφίλ (Bai et al., 2022; DeepSeek-AI, 2024). Η διαπίστωση αυτή υποδηλώνει ότι οι επιλογές κατά την εκπαίδευση ενός μοντέλου επηρεάζουν όχι μόνο τη γενική απόδοσή του αλλά και τον τρόπο με τον οποίο «αντιλαμβάνεται» φαινόμενα όπως η ειρωνεία.

Στο πλαίσιο της ελληνικής ως γλώσσας χαμηλών πόρων, ιδιαίτερη αναφορά αξίζει στη διδακτορική διατριβή του Περήφανου (2019), η οποία αποτελεί την πρώτη εκτεταμένη εμπειρική μελέτη ανάλυσης ελληνικών κειμένων με τεχνικές μηχανικής μάθησης και νευρωνικών γλωσσικών μοντέλων. Στηριζόμενη στην Κατανεμητική Υπόθεση του Harris, σύμφωνα με την οποία σημασιολογικά παρόμοιες λέξεις τείνουν να εμφανίζονται σε παρόμοια περικείμενα, η εργασία χρησιμοποιεί λεξικές ενθέσεις (word embeddings) παραγόμενες από

μοντέλα word2vec, fastText, doc2vec και GloVe σε corpus 325 εκατομμυρίων λέξεων από ελληνόγλωσσα tweets 4.949 χρηστών. Το εύρημα ότι τα μοντέλα αυτά μπορούν να εκπαιδευτούν αποτελεσματικά στην ελληνική (παράγοντας σημασιολογικές αναλογίες τύπου βασιλιάς − άντρας + γυναίκα ≈ βασίλισσα) υποδηλώνει ότι η κατανεμητική φύση της γλώσσας είναι εκμεταλλεύσιμη υπολογιστικά ακόμα και σε γλώσσες με περιορισμένους διαθέσιμους πόρους (Περήφανος, 2019).

Το εύρημα αυτό έχει άμεση συνάφεια με τα ΜΓΜ που εξετάζονται στην παρούσα εργασία. Τα σύγχρονα μοντέλα όπως το Claude και το DeepSeek βασίζονται στην ίδια θεμελιώδη αρχή, την κατανεμητική αναπαράσταση της γλώσσας, αλλά σε πολύ μεγαλύτερη κλίμακα και με αρχιτεκτονική Transformer αντί νευρωνικών δικτύων πρόσθιας τροφοδότησης (Vaswani et al., 2017). Ταυτόχρονα, η διατριβή του Περήφανου (2019) αναδεικνύει ένα κρίσιμο περιορισμό που παραμένει επίκαιρος: η αποτελεσματικότητα των μοντέλων εξαρτάται κρίσιμα από τον όγκο και την ποιότητα των ελληνόγλωσσων εκπαιδευτικών δεδομένων. Σε εφαρμογές που απαιτούν κατανόηση πραγματολογικών φαινομένων, όπως η ειρωνεία, η απλή στατιστική εμφάνιση λέξεων δεν επαρκεί, και τα μοντέλα χρειάζονται εξειδικευμένη έκθεση σε γλωσσικά και πολιτισμικά κατάλληλα δεδομένα (Περήφανος, 2019· Prokopidis & Piperidis, 2020).

Σε ό,τι αφορά γενικότερα τις γλώσσες χαμηλών πόρων, η βιβλιογραφία είναι ενδεικτική των περιορισμών των υπαρχόντων μοντέλων. Οι Conneau et al. (2020) έδειξαν ότι πολυγλωσσικά μοντέλα όπως το XLM-RoBERTa επιτυγχάνουν ικανοποιητική διαγλωσσική μεταφορά (cross-lingual transfer) ακόμα και σε γλώσσες με περιορισμένα εκπαιδευτικά δεδομένα, υπό την προϋπόθεση ότι τα φαινόμενα που ανιχνεύονται έχουν αρκετή δομική ομοιότητα με αυτά των γλωσσών υψηλών πόρων. Η ειρωνεία όμως είναι έντονα πολιτισμικά προσδιορισμένη (Gibbs, 2000), γεγονός που περιορίζει την αποτελεσματικότητα αυτής της μεταφοράς για την ελληνική, ιδίως όταν περιλαμβάνει ιδιωματισμούς, υποκοριστικά με σαρκαστική χροιά και γλωσσικά σχήματα που σπάνια εμφανίζονται σε αγγλόφωνα σώματα κειμένων (Prokopidis & Piperidis, 2020 · Pitenis et al., 2021).

Συνολικά, η βιβλιογραφία αναδεικνύει ότι τα ΜΓΜ αποτελούν υποσχόμενα αλλά όχι ώριμα (ακόμα) εργαλεία για την ανίχνευση ειρωνείας σε γλώσσες χαμηλών πόρων. Η ανάπτυξη εξειδικευμένων ελληνόγλωσσων σωμάτων κειμένων, η συστηματική διερεύνηση

στρατηγικών μηχανικής προτροπών (prompt engineering) και η προσαρμογή (fine-tuning) σε γλωσσικά και πολιτισμικά κατάλληλα δεδομένα αναδεικνύονται ως οι πιο υποσχόμενες κατευθύνσεις για τη βελτίωση της απόδοσης (Devlin et al., 2019; Kojima et al., 2022; Bakagianni et al., 2025).

# 3. Μεθοδολογία και δεδομένα

Το τρίτο κεφάλαιο παρουσιάζει αναλυτικά τη μεθοδολογική σχεδίαση της έρευνας, δηλαδή τις επιλογές και τις διαδικασίες που διέπουν τη συλλογή δεδομένων, την υλοποίηση των συστημάτων και την αξιολόγηση των αποτελεσμάτων. Αρχικά περιγράφεται ο συνολικός ερευνητικός σχεδιασμός και η λογική της συγκριτικής αξιολόγησης (3.1), ενώ στη συνέχεια παρουσιάζονται η σύνθεση και τα χαρακτηριστικά του σώματος των 500 κριτικών, καθώς και η διαδικασία συλλογής τους από τις πλατφόρμες skroutz.gr και bestprice.gr (3.2, 3.2.1). Ακολουθεί η περιγραφή της ανθρώπινης επισημείωσης που αποτελεί την πραγματική κατάσταση (ground truth) της έρευνας, με ειδική αναφορά στη διαδικασία επισημείωσης (annotation) και στους περιορισμούς που απορρέουν από τη χρήση μονού αξιολογητή (3.3, 3.3.1, 3.3.2). Στη συνέχεια περιγράφεται η σχεδίαση και υλοποίηση του αλγορίθμου Python που βασίζεται σε συγκεκριμένους κανόνες (3.4–3.4.2), καθώς και η επιλογή και διαμόρφωση των δύο LLMs, με ιδιαίτερη έμφαση στον σχεδιασμό των προτροπών (3.5–3.5.2). Το κεφάλαιο ολοκληρώνεται με την παρουσίαση των μέτρων αξιολόγησης που χρησιμοποιήθηκαν (3.6) και της μεθοδολογίας ποιοτικής ανάλυσης των αποτελεσμάτων (3.7).

## 3.1 Σχεδιασμός της έρευνας

Η παρούσα εργασία υιοθετεί συγκριτική εμπειρική μεθοδολογία με στόχο την αξιολόγηση τριών διαφορετικών υπολογιστικών προσεγγίσεων ανίχνευσης ειρωνείας σε κείμενα φυσικής γλώσσας: (α) έναν αλγόριθμο βασισμένο σε κανόνες υλοποιημένο σε Python, (β) το μεγάλο γλωσσικό μοντέλο Claude (Anthropic), και (γ) το μεγάλο γλωσσικό μοντέλο DeepSeek (DeepSeek AI). Η αξιολόγηση πραγματοποιείται έναντι ανθρώπινης επισημείωσης που αποτελεί την πραγματική κατάσταση της έρευνας.

Η επιλογή αυτής της τριμερούς σύγκρισης αντιπροσωπεύει τρία θεμελιακά διαφορετικά παραδείγματα επεξεργασίας φυσικής γλώσσας. Η προσέγγιση που βασίζεται σε κανόνες αντανακλά τη λογική των παραδοσιακών συστημάτων ΕΦΓ που βασίζονται σε χειροποίητους κανόνες και λεξιλόγια, ενώ τα ΜΓΜ αντιπροσωπεύουν την τρέχουσα γενιά νευρωνικών μοντέλων που μαθαίνουν στατιστικές αναπαραστάσεις από τεράστια σώματα κειμένων. Η σύγκρισή τους επιτρέπει τον εντοπισμό τόσο των δυνατοτήτων όσο και των ορίων

κάθε παραδείγματος για ένα έργο που απαιτεί πραγματολογική κατανόηση εις βάθος (Grice, 1975).

## 3.2 Δημιουργία σώματος κειμένων

Για τη συλλογή των κριτικών και την συγκρότηση του υπό ανάλυση σώματος κειμένων, επιστρατεύτηκαν διαδικτυακές κριτικές χρηστών. Οι κριτικές αντλήθηκαν από τους διαδικτυακούς τόπους skroutz.gr και bestprice.gr. Οι κριτικές αφορούν ποικίλες κατηγορίες προϊόντων, συμπεριλαμβανομένων τεχνολογικών συσκευών, παπουτσιών/ρούχων και οικιακών συσκευών, με στόχο να αντιπροσωπευτεί ένα ευρύ φάσμα γλωσσικών φαινομένων και θεματικών πεδίων. Κατά τη συγκρότηση του σώματος κειμένων, επιλέχθηκαν 500 διαφορετικές κριτικές, οι οποίες περιλάμβαναν φράσεις με ειρωνεία, με αρκετά άλλα σχήματα λόγου, αλλά και φράσεις που χαρακτηρίζονται εντελώς κυριολεκτικές. Η επιλογή έγινε με βάση τα γλωσσικά στοιχεία που υπάρχουν στις προτάσεις. Επιδιώχθηκε η συγκρότηση ενός σώματος που να ανταποκρίνεται στις ανάγκες της έρευνας. Υπάρχουν προτάσεις με δύσκολα ανιχνεύσιμες και «κρυμμένες» ειρωνείες, ώστε να ελεγχθούν οι ικανότητες των μοντέλων. Η επιλογή δεν ήταν τυχαία, καθώς σε τέτοια περίπτωση θα απουσίαζε η ποικιλομορφία και τα παραδείγματα δεν θα ανταποκρινόταν στην πραγματικότητα.

Η επιλογή κριτικών καταναλωτών ως πηγή κειμένων υπαγορεύτηκε από τη θεωρητική παραδοχή ότι αυτό το είδος γραπτού λόγου παρουσιάζει ιδιαίτερα υψηλή συχνότητα εμφάνισης ειρωνείας σε φυσικό, μη επιμελημένο πλαίσιο (Maynard & Greenwood, 2014). Οι χρήστες πλατφορμών αξιολόγησης προϊόντων χρησιμοποιούν συχνά σαρκασμό για να εκφράσουν δυσαρέσκεια, υπερβολή για να τονίσουν θετικές ιδιότητες, και ιδιωματικές μεταφορές του καθημερινού ελληνικού λόγου για να περιγράψουν την εμπειρία τους. Κριτήριο επιλογής αποτέλεσε η γλωσσική πολυπλοκότητα και η ποικιλομορφία που θα μπορούσε να συμβάλλει στην εξαγωγή συμπερασμάτων.

Στη συνέχεια, οι κριτικές αποθηκεύτηκαν, δίχως όμως να υπάρχει μεταβολή του περιεχομένου. Μετατράπηκαν σε μια μηχαναγνώσιμη μορφή χωρίς να αλλαχθούν ως προς την δομή και το περιεχόμενο. Διατηρήθηκαν όλα τα σημεία στίξης, τα ορθογραφικά λάθη, οι εμβόλιμες ξένες λέξεις, τα greeklish και η δομή.

## 3.3. Διαδικασία επισημείωσης

Η ανθρώπινη επισημείωση πραγματοποιήθηκε από έναν αξιολογητή. Κάθε κείμενο αξιολογήθηκε ανεξάρτητα ως προς τη δυαδική ερώτηση: περιέχει ειρωνεία; με δυνατές αποκρίσεις ΝΑΙ/ΟΧΙ. Όπου κρίθηκε απαραίτητο, καταγράφηκε σύντομη δικαιολόγηση της απόφασης με επισήμανση της συγκεκριμένης λέξης ή φράσης που αιτιολογεί την κατηγοριοποίηση.

Για τους σκοπούς της επισημείωσης υιοθετήθηκαν οι ορισμοί που αναφέρθηκαν και παραπάνω. Ως ειρωνεία ορίστηκε κάθε περίπτωση κατά την οποία ο ομιλητής εκφέρει το αντίθετο ή διαφορετικό από αυτό που κυριολεκτικά δηλώνουν τα λόγια του, με σκοπό να εκφράσει κριτική, δυσαρέσκεια ή χιούμορ (Sperber & Wilson, 1981), σε συνδυασμό με τα πραγματικά κοινωνικά και γλωσσολογικά περικείμενα.

Πρέπει να επισημανθεί ότι η επισημείωση πραγματοποιήθηκε από έναν μόνο αξιολογητή, γεγονός που συνιστά μεθοδολογικό περιορισμό. Η ειρωνεία είναι υποκειμενικό φαινόμενο. Ειδικότερα, έρευνες έχουν δείξει ότι ακόμα και μεταξύ ανθρώπων η συμφωνία κυμαίνεται τυπικά στο κ = 0.40–0.65 (Artstein & Poesio, 2008· Van Hee et al., 2018). Τα αποτελέσματα πρέπει να ερμηνεύονται λαμβάνοντας υπόψη αυτόν τον περιορισμό και η επέκταση της επαλήθευσης σε δεύτερο αξιολογητή αποτελεί σαφή προτεραιότητα για μελλοντική έρευνα.

## 3.4 Σχεδιασμός και υλοποίηση του αλγόριθμου

Ο κώδικας που χρησιμοποιήθηκε για την ταξινόμηση αναπτύχθηκε από τον Κωνσταντίνο Περήφανο και παραχωρήθηκε για τους σκοπούς της παρούσας εργασίας (Περήφανος, 2026). Το script υλοποιείται σε Python 3 και αξιοποιεί τη βιβλιοθήκη instructor σε συνδυασμό με το Anthropic API, επιτρέποντας δομημένη εξαγωγή αποτελεσμάτων μέσω Pydantic μοντέλων. Συγκεκριμένα, για κάθε κείμενο το σύστημα επιστρέφει τρία πεδία: την ετικέτα ταξινόμησης (irony / not_irony), βαθμό βεβαιότητας (confidence score, 0.0–1.0) και σύντομη αιτιολόγηση της απόφασης (reasoning). Το prompt σχεδιάστηκε ώστε να κατευθύνει το μοντέλο να εξετάζει τρεις τύπους ειρωνείας: λεκτική ειρωνεία, σαρκασμό και καταστασιακή ειρωνεία. Ο πλήρης κώδικας παρατίθεται στο Παράρτημα.

Ο αλγόριθμος υλοποιήθηκε σε Python 3 με αποκλειστική χρήση βιβλιοθηκών ανοικτού κώδικα: pandas για τη διαχείριση των δεδομένων, re για την εφαρμογή κανονικών εκφράσεων (regular expressions), matplotlib και openpyxl για την οπτικοποίηση και εξαγωγή αποτελεσμάτων. Το σώμα κειμένων των 500 κριτικών φορτώθηκε από αρχείο Excel (.xlsx) μέσω της βιβλιοθήκης openpyxl και επεξεργάστηκε διαδοχικά, με το σύστημα να εκτελεί σε κάθε κείμενο τον πλήρη κύκλο ανίχνευσης.

Η ανίχνευση ειρωνείας βασίζεται σε δύο συμπληρωματικούς μηχανισμούς. Ο πρώτος αφορά λεξιλογικά και ορθογραφικά σήματα: ανιχνεύονται μέσω δομικών σχημάτων κανονικών εκφράσεων (regex) όπως η παρουσία αποσιωπητικών (…), πολλαπλών θαυμαστικών (!!), και φράσεων με υπονοούμενη ειρωνική χρήση (*βέβαια, ναι ναι, μόνο Χ ώρες*). Ο δεύτερος μηχανισμός εντοπίζει σημασιολογική αντίθεση: αν το ίδιο κείμενο περιέχει ταυτόχρονα θετικές λέξεις (π.χ. *άριστο, εξαιρετικό, τέλειο*) και αρνητικές (π.χ. *πρόβλημα, καθυστέρηση, απαράδεκτο*), το σύστημα προσθέτει επιπλέον βαρύτητα στο σκορ ειρωνείας. Η τελική ταξινόμηση γίνεται με κατώφλι: οποιοδήποτε σκορ ≥ 1 οδηγεί σε ετικέτα «ΝΑΙ». Αξίζει να σημειωθεί ότι η διαδικασία είναι πλήρως ντετερμινιστική: για το ίδιο κείμενο το σύστημα παράγει πάντα την ίδια απόφαση, χωρίς καμία στοχαστική συνιστώσα — σε αντίθεση με τα ΜΓΜ, των οποίων η έξοδος επηρεάζεται από παραμέτρους δειγματοληψίας (Ouyang et al., 2022).

Τα αποτελέσματα αποθηκεύτηκαν σε αρχείο Excel πολλαπλών φύλλων, που περιλαμβάνει το σύνολο των κριτικών με τις ετικέτες και τα επιμέρους σκορ. Η δομή αυτή επέτρεψε την άμεση σύγκριση με τα αποτελέσματα των ΜΓΜ και τον υπολογισμό όλων των μετρικών αξιολόγησης που παρουσιάζονται στην ενότητα 4.

## 3.5. Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ)

### 3.5.1 Επιλογή Μοντέλων

Για την έρευνα επιλέχθηκαν δύο ΜΓΜ: το Claude (Anthropic) και το DeepSeek (DeepSeek AI). Η επιλογή τους υπαγορεύτηκε από τη διαφορετική αρχιτεκτονική καταγωγή και εκπαίδευσή τους: το Claude έχει αναπτυχθεί από αμερικανική εταιρεία με έμφαση στο Constitutional AI και το RLHF (Bai et al., 2022), ενώ το DeepSeek προέρχεται από κινεζική

ερευνητική ομάδα με διαφορετική προσέγγιση εκπαίδευσης (DeepSeek-AI, 2024). Αυτή η διαφορετικότητα καθιστά τη σύγκρισή τους ερευνητικά ενδιαφέρουσα, καθώς πιθανές διαφορές στα αποτελέσματα μπορούν να αποδοθούν σε διαφορές στα δεδομένα εκπαίδευσης, στη μεθοδολογία εκπαίδευσης ή στον στόχο.

Και τα δύο μοντέλα χρησιμοποιήθηκαν στην εκμάθηση με μηδενικά παραδείγματα (zero-shot) εκδοχή τους, δηλαδή χωρίς παροχή παραδειγμάτων (few-shot examples) στην προτροπή, ώστε να αξιολογηθεί η εγγενής τους ικανότητα κατανόησης της μη κυριολεκτικής γλώσσας χωρίς ειδική καθοδήγηση. Η αξιολόγηση με διαδικασία εκμάθησης με μηδενικά παραδείγματα (zero-shot) αποτελεί μεθοδολογική επιλογή που επιτρέπει τη σύγκριση με το σύστημα που βασίζεται σε κανόνες σε κοινή βάση αφετηρίας.

### 3.5.2 Σχεδιασμός προτροπών

Κάθε κείμενο υποβλήθηκε στα ΜΓΜ μέσω μιας ενιαίας, σταθερής προτροπής που παρέμεινε αμετάβλητη για όλα τα 500 κείμενα και για τα δύο μοντέλα (Wei et al., 2022).

Η προτροπή ζητούσε από κάθε μοντέλο να αποφανθεί για κάθε κείμενο ως προς ένα ερώτημα: εντοπίζεται ειρωνεία; με μονολεκτική απάντηση ΝΑΙ/ΟΧΙ. Στη συνέχεια ζητήθηκε με νέα οδηγία να παράσχει σύντομη δικαιολόγηση με αναφορά στη συγκεκριμένη λέξη ή φράση που τεκμηριώνει τη σκέψη του. Η απαίτηση δικαιολόγησης εξυπηρετεί διπλό σκοπό: αφενός επιτρέπει την ποιοτική ανάλυση της συλλογιστικής των μοντέλων, αφετέρου έχει τεκμηριωθεί ότι η παραγωγή εξήγησης πριν από την τελική απόφαση βελτιώνει την ακρίβεια σε «ασκήσεις συλλογιστικής» (reasoning tasks) (Kojima et al., 2022).

Το ακριβές κείμενο της προτροπής ήταν το εξής:

> «Θα ήθελα παρακαλώ να διαβάσεις το αρχείο με τις προτάσεις / κριτικές. Να μου πεις σε μια - μια ξεχωριστά αν υπάρχει ειρωνεία ή όχι. Παρακαλώ να απαντήσεις μόνο αυτό δίχως αιτιολόγηση με τη μορφή π.χ.:
> 1. Ειρωνεία
> 1. Όχι Ειρωνεία

Το δεύτερο κείμενο της οδηγίας που αποσκοπούσε στην δικαιολόγηση της απάντησης στη συνέχεια ήταν το εξής:

> «Ευχαριστώ. Τώρα θα ήθελα να μου επεξηγήσεις γιατί έκανες κάθε μια από αυτές τις επιλογές. π.χ.: στο 3. υπάρχει ειρωνεία λόγω της φράσης .....»

Θα πρέπει να σημειωθεί ότι τα ΜΓΜ παρουσιάζουν μη ντετερμινιστική συμπεριφορά λόγω της στοχαστικής φύσης της παραγωγής κειμένου. Στην παρούσα έρευνα κάθε κείμενο υποβλήθηκε μία φορά σε κάθε μοντέλο, χωρίς επανάληψη. Αυτό αποτελεί έναν σημαντικό περιορισμό που πρέπει να ληφθεί υπόψη κατά την ερμηνεία των αποτελεσμάτων (Ouyang et al., 2022).

## 3.6 Μέτρα Αξιολόγησης

Για την αξιολόγηση της απόδοσης κάθε συστήματος χρησιμοποιήθηκε συνδυασμός μετρικών, καθώς καμία μεμονωμένη μετρική δεν επαρκεί για πλήρη εικόνα σε ανισόρροπα σύνολα δεδομένων (datasets).

Ως προς τα μέτρα αξιοποιήθηκαν οι εξής τιμές:

- Σωστά Θετικά / TP
- Λανθασμένα Θετικά / FP
- Λανθασμένα Αρνητικά / FN
- Σωστά Αρνητικά / TN
- Ακρίβεια Θετικών / Precision
- Ανάκληση / Recall
- F1
- Ακρίβεια / Accuracy
- Ισορροπημένη Ακρίβεια / Balanced Accuracy
- Δείκτης Συμφωνίας κ (έναντι Πραγματικής Κατάστασης) / Cohen's κ (vs GT)

Οι μετρικές Ακρίβεια Θετικών, Ανάκληση και F1 αποτελούν τις βασικές μετρικές αξιολόγησης δυαδικής ταξινόμησης. Η Ακρίβεια Θετικών εκφράζει το ποσοστό των θετικών προβλέψεων

που είναι σωστές, η Ανάκληση εκφράζει το ποσοστό των πραγματικών θετικών περιπτώσεων που εντοπίστηκαν, και το F1 αποτελεί το αρμονικό μέσο τους (Manning et al., 2008).

Η παράμετρος της Συνολικής Ακρίβειας (Overall Accuracy) αναφέρεται για λόγους πληρότητας, αλλά δεν αποτελεί αξιόπιστο δείκτη στο παρόν σώμα κειμένων λόγω κλασικής ανισορροπίας κλάσεων (class imbalance). Χαρακτηριστικό παράδειγμα η περίπτωση όπου ένα σύστημα που απαντά πάντα «ΟΧΙ» θα επιτύγχανε ένα πολύ υψηλό ποσοστό στην μέτρηση της Ακρίβειας χωρίς καμία πρακτική απόδοση (Sokolova & Lapalme, 2009). Η μετρική Ισορροπημένη Ακρίβεια (Balanced Accuracy) υπολογίζει τον μέσο όρο της Ανάκλησης (Recall) ανά κλάση, αντιμετωπίζοντας την ανισορροπία πιο αξιόπιστα από την συνολική ακρίβεια.

Η παράμετρος της Συμφωνίας κ (Cohen's Kappa) μετρά τη συμφωνία μεταξύ δύο ταξινομητών διορθώνοντας για τη συμφωνία που αναμένεται από τύχη. Αποτελεί ιδιαίτερα σημαντική μετρική στην παρούσα έρευνα διότι επιτρέπει άμεση σύγκριση τόσο μεταξύ κάθε μοντέλου και της Πραγματικής Κατάστασης όσο και μεταξύ των μοντέλων μεταξύ τους (inter-model agreement). Ο οδηγός ερμηνείας είναι ο ακόλουθος: $\kappa < 0.20$ = ελάχιστη συμφωνία, 0.21–0.40 = μέτρια, 0.41–0.60 = μέτρια προς ικανοποιητική, 0.61–0.80 = ισχυρή, $> 0.80$ = σχεδόν τέλεια συμφωνία (Landis & Koch, 1977).

## 3.7 Ποιοτική ανάλυση

Παράλληλα με την ποσοτική αξιολόγηση, πραγματοποιήθηκε ποιοτική ανάλυση σφαλμάτων με στόχο την εξήγηση των μοτίβων λαθών κάθε συστήματος. Ενώ τα ποσοτικά μέτρα (F1, Precision, Recall, Cohen's κ) αποτυπώνουν το πόσο καλά αποδίδει ένα σύστημα συνολικά, δεν μπορούν να αποκαλύψουν γιατί αποτυγχάνει σε συγκεκριμένες περιπτώσεις και ποια γλωσσολογικά χαρακτηριστικά του κειμένου ευθύνονται για τα λάθη (Krippendorff, 2019).

Η ανάλυση επικεντρώθηκε σε τρεις άξονες. Αρχικά, εξετάστηκαν συστηματικά οι περιπτώσεις ψευδώς θετικών (False Positive) κάθε συστήματος, δηλαδή κριτικές που χαρακτηρίστηκαν ως ειρωνικές ενώ δεν ήταν, με στόχο να εντοπιστούν τα γλωσσολογικά στοιχεία που «παραπλανούν» κάθε σύστημα. Στη συνέχεια, εξετάστηκαν οι περιπτώσεις ψευδώς αρνητικών (False Negative), δηλαδή κριτικές που δεν εντοπίστηκαν από κανένα

σύστημα, έτσι ώστε να αναδειχθούν οι τύποι ειρωνείας που παραμένουν αόρατοι. Τέλος, αναλύθηκαν οι δικαιολογήσεις που παρείχαν τα ΜΓΜ για κάθε απόφασή τους, ως ποιοτική ένδειξη της εσωτερικής συλλογιστικής τους: το Claude και το DeepSeek δεν επιστρέφουν μόνο ετικέτα αλλά και αιτιολόγηση, γεγονός που επιτρέπει να εξεταστεί αν η απόφαση βασίζεται σε γλωσσολογικά έγκυρα κριτήρια ή σε επιφανειακά μοτίβα (Oprea & Bara, 2025).

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι περιπτώσεις όπου τα τρία συστήματα διαφωνούν μεταξύ τους, καθώς αυτές αποκαλύπτουν τις «γκρίζες ζώνες» του φαινομένου: κείμενα που δεν φέρουν ρητά λεξιλογικά σήματα ειρωνείας αλλά απαιτούν πραγματολογική συλλογιστική για την ερμηνεία τους. Οι περιπτώσεις αυτές είναι ακριβώς εκείνες που αναδεικνύουν τα θεωρητικά όρια κάθε προσέγγισης και τεκμηριώνουν εμπειρικά το επιχείρημα ότι η ειρωνεία λειτουργεί σε επίπεδο που υπερβαίνει τη λεξιλογική επιφάνεια του κειμένου (Attardo, 2000; Grice, 1975).

# 4. Αποτελέσματα

Το τέταρτο κεφάλαιο παρουσιάζει και ερμηνεύει τα αποτελέσματα της συγκριτικής αξιολόγησης των τριών συστημάτων. Ξεκινά με μια πρώτη ανάλυση και παρουσίαση του συνόλου των αποτελεσμάτων (4.1, 4.1.1), ενώ στη συνέχεια παρέχεται η γενική εικόνα της κύριας ανάλυσης έναντι της πραγματικής κατάστασης (4.2). Ακολουθεί αναλυτική παρουσίαση των αποτελεσμάτων ανά σύστημα με εστίαση στην ειρωνεία (4.3, 4.3.1) και επιστημονική ερμηνεία των ευρημάτων, αναδεικνύοντας τι αποκαλύπτει η σύγκριση των τριών προσεγγίσεων σε θεωρητικό επίπεδο (4.4). Το κεφάλαιο συνεχίζει με ποιοτική ανάλυση χαρακτηριστικών περιπτώσεων λαθών (4.5) και εξέταση των μοτίβων δικαιολόγησης που εμφανίζουν το Claude και το DeepSeek (4.5.1, 4.5.2). Ολοκληρώνεται με κριτική αποτίμηση των περιορισμών της έρευνας και των ορίων που θέτουν τα δεδομένα στην ερμηνεία των αποτελεσμάτων (4.6).

## 4.1 Πρώτη Ανάλυση

Συγκρίνονται τρεις ποιοτικά διαφορετικές προσεγγίσεις ανίχνευσης ειρωνείας σε 500 ελληνικές κριτικές από το διαδίκτυο. Έχουμε έναν αλγόριθμο της Python που βασίζεται σε κανόνες (rule-based), και δύο μεγάλα γλωσσικά μοντέλα, το Claude και το DeepSeek. Πρόκειται για τριμερή σύγκριση ισότιμων αλλά αρκετά διαφορετικών στρατηγικών επεξεργασίας φυσικής γλώσσας.

Το κεντρικό εύρημα είναι σαφές: οι τρεις προσεγγίσεις διαφωνούν συστηματικά μεταξύ τους, με τρόπους που δεν είναι τυχαίοι αλλά αποκαλύπτουν θεμελιακές διαφορές στο τι «βλέπει» η κάθε μία ως ειρωνεία.

Μια πρώτη, ενδεικτική εικόνα της συμπεριφοράς κάθε συστήματος προκύπτει απλά από τον αριθμό των θετικών ανιχνεύσεων (Πίνακας 1). Ο ανθρώπινος αξιολογητής χαρακτήρισε 194 από τις 500 κριτικές ως ειρωνικές (38,8%). Ο αλγόριθμος Python κινήθηκε πλησιέστερα σε αυτή την αναλογία με 185 θετικές ανιχνεύσεις (37%), υποδηλώνοντας ότι η συνολική κατανομή των προβλέψεών του είναι ευθυγραμμισμένη με την πραγματική κατάσταση. Το Claude υπερεκτίμησε σημαντικά τη συχνότητα της ειρωνείας με 259 θετικές ανιχνεύσεις (51.8%), σχεδόν μία στις δύο κριτικές κρίθηκε ειρωνική, αποκαλύπτοντας την τάση υπεργενίκευσης που επιβεβαιώνεται στη συνέχεια από τα 141 FP. Το DeepSeek κινήθηκε στον απόλυτο αντίποδα με μόλις 49 θετικές ανιχνεύσεις (9.8%), δηλαδή απέρριψε ως μη-

ειρωνικές τις 9 στις 10 κριτικές, αντικατοπτρίζοντας την εξαιρετικά συντηρητική στρατηγική που το χαρακτηρίζει.

| **Προσέγγιση** | **Αριθμός Ειρωνειών** |
|---|---|
| Ανθρώπινος Αξιολογητής | 194/500 |
| Αλγόριθμος Python | 185/500 |
| Claude | 259/500 |
| DeepSeek | 49/500 |

**Πίνακας 1: Πόσες φορές ανιχνεύει θετικά κάθε προσέγγιση (n=500)**

**Τριπλή συμφωνία (και τα 3 ίδια απάντηση):**

191/500 Και τα 3 σωστά (vs GT): 160/500 (32%): Εξίσου αποκαλυπτικά είναι αυτά τα δεδομένα συμφωνίας όπου και τα τρία συστήματα έδωσαν την ίδια απάντηση σε 191 από τις 500 κριτικές (38.2%), αλλά από αυτές μόνο οι 160 (32%) ήταν σωστές έναντι της πραγματικής κατάστασης. Αυτό σημαίνει ότι ακόμα και στις περιπτώσεις όπου τα τρία συστήματα συμφωνούν, η ομοφωνία τους δεν αποτελεί εγγύηση ορθής ταξινόμησης. Το εύρημα αυτό, αμφισβητεί τη διαισθητική παραδοχή ότι η συναίνεση πολλαπλών συστημάτων αυξάνει την αξιοπιστία (Polikar, 2006). Η παρατήρηση αυτή ενισχύει το επιχείρημα ότι τα τρία συστήματα μοιράζονται «τυφλές γωνίες», δηλαδή κατηγορίες ειρωνείας που κανένα δεν εντοπίζει αξιόπιστα, πιθανώς λόγω της απουσίας πραγματολογικού περικειμένου και εξωγλωσσικής πληροφορίας (Attardo, 2000; Grice, 1975).

| **Ζεύγος** | **Ειρωνεία κ** |
|---|---|
| GT ↔ Python | +0.495 |
| GT ↔ Claude | +0.139 |
| GT ↔ DeepSeek | +0.224 |
| Claude ↔ DeepSeek | +0.044 |

| | | |
|---|---|---|
| Claude ↔ Python | +0.120 | |
| DeepSeek ↔ Python | +0.241 | |
| GT ↔ Claude | −0.072 | −0.020 |

**Πίνακας 2: Συμφωνία ανά ζεύγος (Cohen's kappa)**

## 4.2 Ανάλυση με ανθρώπινο αξιολογητή και πραγματική κατάσταση

Η κύρια ανάλυση βασίζεται στη σύγκριση των προβλέψεων κάθε συστήματος έναντι της ανθρώπινης επισημείωσης στο σύνολο των 500 κριτικών. Από τις 500 κριτικές, ο ανθρώπινος αξιολογητής χαρακτήρισε 194 ως ειρωνικές (38.8%) και 306 ως μη ειρωνικές (61.2%), διαμορφώνοντας ένα μέτρια ανισόρροπο σύνολο δεδομένων που επιβάλλει την αξιολόγηση μέσω μετρικών πέραν της απλής ακρίβειας, όπως F1, η Ισορροπημένη Ακρίβεια και ο δείκτης κ (Sokolova & Lapalme, 2009).

Σε επίπεδο συνολικής κατανομής προβλέψεων, τα τρία συστήματα εμφανίζουν εντυπωσιακά διαφορετική συμπεριφορά. Ο αλγόριθμος Python ανίχνευσε 185 κριτικές ως ειρωνικές (37%), παραμένοντας πλησιέστερα στην πραγματική κατανομή. Το Claude ανίχνευσε 259 (51.8%), υπερεκτιμώντας σχεδόν κατά 33% τη συχνότητα της ειρωνείας. Το DeepSeek ανίχνευσε μόλις 49 (9.8%), υποεκτιμώντας δραματικά το φαινόμενο. Η απόκλιση αυτή δεν είναι απλώς στατιστική, αλλά, αντίθετα, αντικατοπτρίζει θεμελιακά διαφορετικές «θεωρίες» για το τι συνιστά ειρωνεία που έχει αναπτύξει κάθε σύστημα (Van Hee et al., 2018).

Σε επίπεδο συμφωνίας με την πραγματική κατάσταση, ο αλγόριθμος Python επιτυγχάνει F1 = 0.686 και Δείκτης κ = 0.495, τιμή που εντάσσεται στη ζώνη μέτριας προς ικανοποιητικής συμφωνίας (Landis & Koch, 1977). Το Claude ακολουθεί με F1 = 0.521 και κ = 0.139, ενώ το DeepSeek εμφανίζει το χαμηλότερο F1 = 0.346 παρά την υψηλή Ακρίβεια (0.857), λόγω της εξαιρετικά χαμηλής Ανάκλησης (0.216). Αξίζει να σημειωθεί ότι καμία από τις τρεις τιμές κ δεν υπερβαίνει το 0.50, δηλαδή κανένα σύστημα δεν επιτυγχάνει ικανοποιητική συμφωνία με τον ανθρώπινο αξιολογητή κατά την κλασική κλίμακα ερμηνείας των Artstein και Poesio (2008). Αυτό δεν αποτελεί τυχαίο εύρημα της εργασίας, αλλά

αποτυπώνει μια ευρύτερη πρόκληση που έχει καταγραφεί επανειλημμένα στη βιβλιογραφία για την ανίχνευση ειρωνείας σε γλώσσες χαμηλών πόρων (Pitenis et al., 2021; Prokopidis & Piperidis, 2020).

Σε επίπεδο διασυστημικής συμφωνίας, τα αποτελέσματα είναι εξίσου ενδιαφέροντα. Η συμφωνία Claude↔DeepSeek (κ = 0.044) είναι ουσιαστικά μηδενική, υποδηλώνοντας ότι τα δύο μοντέλα δεν προσεγγίζουν το ίδιο «πρόβλημα» με τον ίδιο τρόπο. Η συμφωνία Python↔Claude και Python↔DeepSeek είναι επίσης χαμηλή, επιβεβαιώνοντας ότι τα τρία συστήματα «κοιτούν» διαφορετικά πράγματα στο κείμενο. Το εύρημα αυτό έχει σημαντικές συνέπειες: αναδεικνύει ότι η ειρωνεία δεν είναι ένα αντικειμενικό, μονοσήμαντο χαρακτηριστικό του κειμένου, αλλά ένα πολυεπίπεδο φαινόμενο που ορίζεται διαφορετικά ανάλογα με τον παρατηρητή ανθρώπινο ή αλγοριθμικό (Gibbs, 2000; Grice, 1975).

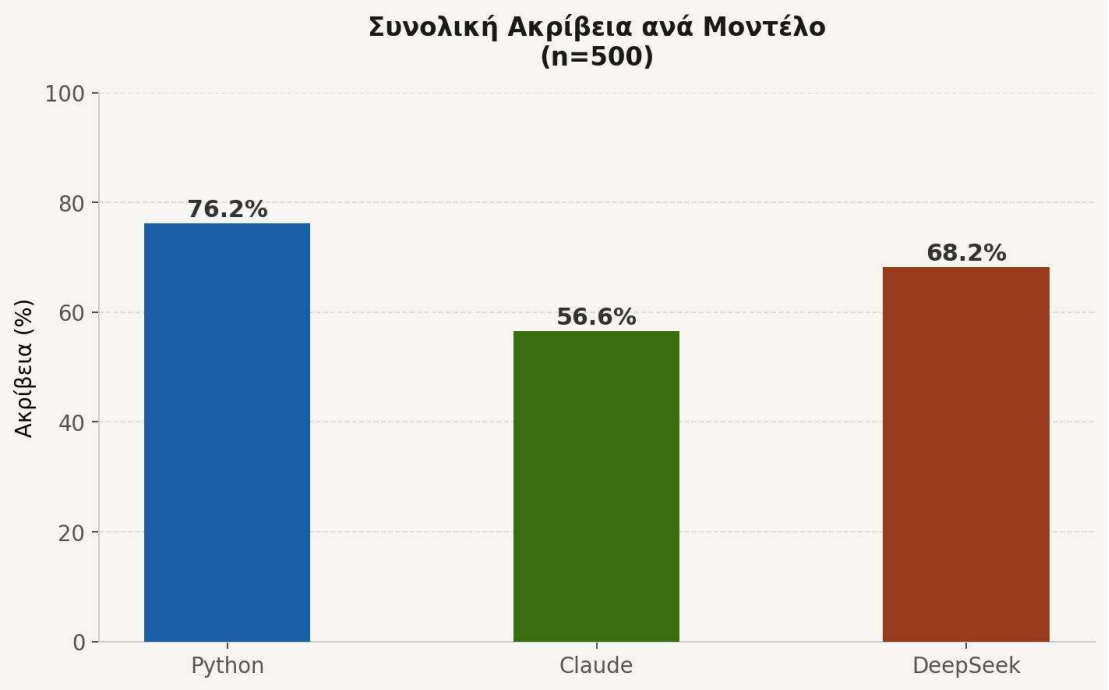
## 4.3 Αποτελέσματα ανά Σύστημα

Στη συνέχεια, παρουσιάζονται αναλυτικά τα αποτελέσματα του κάθε συστήματος ξεχωριστά, βάσει του συνόλου των μετρικών αξιολόγησης που περιγράφηκαν στην ενότητα 3.6. Η ανάλυση οργανώνεται γύρω από τέσσερις βασικές διαστάσεις: την ικανότητα κάθε συστήματος να εντοπίζει πραγματικές ειρωνικές περιπτώσεις (Recall), την αξιοπιστία των θετικών του αποφάσεων (Precision), τη συνολική ισορροπία μεταξύ των δύο (F1), και τον βαθμό συμφωνίας με τον ανθρώπινο αξιολογητή (Cohen's κ). Ο Πίνακας 3 παρουσιάζει τα πλήρη αποτελέσματα για κάθε σύστημα, συμπεριλαμβανομένων των απόλυτων αριθμών Αληθώς και Ψευδών Θετικών και Αρνητικών, οι οποίοι επιτρέπουν μια πιο άμεση και διαισθητική κατανόηση του τύπου των σφαλμάτων κάθε προσέγγισης.

| **Μέτρο** | **Python** | **Claude** | **DeepSeek** |
|---|---|---|---|
| TP / Αληθώς Θετικά | 130 | 118 | 42 |
| FP / Ψευδώς Θετικά | 55 | 141 | 7 |
| FN / Ψευδώς Αρνητικά | 64 | 76 | 152 |

| **Μέτρο** | **Python** | **Claude** | **DeepSeek** |
|---|---|---|---|
| TN / Ψευδώς Αρνητικά | 251 | 165 | 299 |
| Precision / Ακρίβεια Θετικών | 0.703 | 0.456 | 0.857 |
| Recall / Ανάκληση | 0.670 | 0.608 | 0.216 |
| F1 | 0.686 | 0.521 | 0.346 |
| Accuracy / Ακρίβεια | 0.762 | 0.566 | 0.682 |
| Balanced Accuracy / Ισορροπημένη Ακρίβεια | 0.745 | 0.574 | 0.597 |
| Cohen's κ (vs GT) / Δείκτης Συμφωνίας Cohen's κ (έναντι Πραγματικής Κατάστασης) | +0.495 | +0.139 | +0.224 |

**Πίνακας 3: Πλήρη αποτελέσματα για το κάθε σύστημα**




**Εικόνα 2: Ακρίβεια ανίχνευσης ανά μοντέλο**

Τα αποτελέσματα της παρούσας έρευνας θέτουν υπό αμφισβήτηση μια μερίδα ερευνών που υποστηρίζουν ότι τα ΜΓΜ, κατά βάση, υπερτερούν των προσεγγίσεων που βασίζονται σε κανόνες σε εργασίες κατανόησης της μη κυριολεκτικής γλώσσας (Brown et al., 2020). Στο παρόν σώμα κειμένων, ο αλγόριθμος Python επιτυγχάνει την καλύτερη συνολική ισορροπία (F1 = 0.686), ενώ κανένα από τα δύο ΜΓΜ δεν πλησιάζει αυτή την επίδοση. Το εύρημα αυτό δεν είναι απαραίτητα έκπληξη αν ληφθεί υπόψη η φύση του σώματος κειμένων: οι κριτικές καταναλωτών σε πλατφόρμες όπως το skroutz.gr τείνουν να χρησιμοποιούν σχετικά ρητά και επαναλαμβανόμενα ειρωνικά σχήματα τα οποία ευνοούν ακριβώς τη λεξιλογική λογική ενός συστήματος βασισμένου σε συγκεκριμένους κανόνες (Riloff et al., 2013; Maynard & Greenwood, 2014).

Ωστόσο, η υπεροχή του αλγορίθμου είναι εύθραυστη και περιορισμένη. Ο αλγόριθμος αποτυγχάνει συστηματικά σε ειρωνεία που δεν συνοδεύεται από τυπικούς γλωσσικούς δείκτες, έμμεση ειρωνεία, αποστασιοποίηση, σαρκαστικές ετικέτες, δηλαδή σε ακριβώς εκείνες τις περιπτώσεις που απαιτούν πραγματολογική κατανόηση (Attardo, 2000; Grice, 1975). Αντίθετα, το Claude εντοπίζει με επιτυχία αυτές τις δυσκολότερες περιπτώσεις, επιβεβαιώνοντας ότι τα ΜΓΜ διαθέτουν ανώτερη ικανότητα πραγματολογικής συλλογιστικής,

αλλά με κόστος: 141 ψευδώς θετικές ταξινομήσεις αποκαλύπτουν ότι το μοντέλο υπεργενικεύει, ερμηνεύοντας ως ειρωνικές δομές που απλώς είναι αντιθετικές ή έντονες (Oprea, 2025).

Η συμπεριφορά του DeepSeek παρουσιάζει ιδιαίτερο θεωρητικό ενδιαφέρον. Η εξαιρετικά υψηλή Ακρίβεια (0.857) με ταυτόχρονα χαμηλό Ανάκληση (0.216) υποδηλώνει ότι το μοντέλο λειτουργεί με πολύ υψηλό επίπεδο βεβαιότητας: αποφαίνεται για ειρωνεία μόνο όταν τα σήματα είναι εξαιρετικά σαφή, αποφεύγοντας να προκρίνει ως ειρωνεία κάθε ασαφή ή διφορούμενη περίπτωση (Jang & Shin, 2020). Αυτή η στρατηγική μειώνει δραστικά τα ψευδώς θετικά, αλλά παράλληλα το μοντέλο, παρουσιάζεται «τυφλό» στην πλειονότητα των ειρωνικών κειμένων, οδηγώντας στο χαμηλότερο F1 (0.346) από τα τρία συστήματα. Η διαφορά στη συμπεριφορά Claude-DeepSeek πιθανώς αντικατοπτρίζει διαφορές στα δεδομένα εκπαίδευσης και στις στρατηγικές RLHF των δύο μοντέλων (Bai et al., 2022; DeepSeek-AI, 2024), καθώς και το γεγονός ότι η ελληνική γλώσσα υποεκπροσωπείται στα σώματα κειμένων που χρησιμοποιούνται για την προ εκπαίδευση (Pitenis et al., 2021).

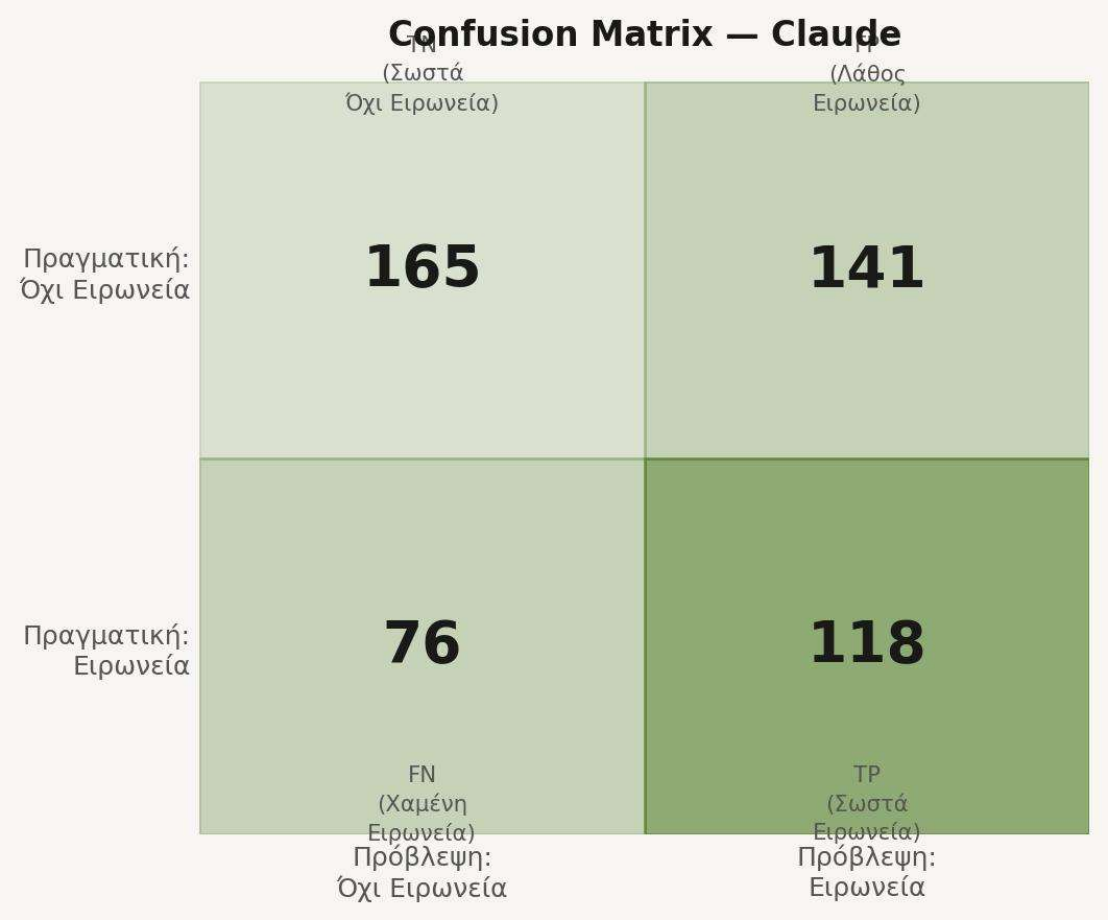
Κεντρικό θεωρητικό συμπέρασμα της κριτικής ανάλυσης αποτελεί η διαπίστωση ότι η σχεδόν μηδενική συμφωνία μεταξύ Claude και DeepSeek (κ = 0.044) δεν αντανακλά απλώς διαφορετική απόδοση, αλλά δομική διαφορά. Το εύρημα αυτό συνάδει με την παρατήρηση των Ortega Bueno (2025) και Van Hee et al. (2018) ότι η ειρωνεία δεν είναι μία ενιαία κατηγορία αλλά ένα φάσμα φαινομένων που διαφορετικά συστήματα, σαφώς και διαφορετικοί άνθρωποι, οριοθετούν με διαφορετικό τρόπο. Η απουσία δεύτερου ανθρώπινου αξιολογητή δεν επιτρέπει να αποφανθούμε με βεβαιότητα ποιο σύστημα «έχει δίκιο» στις αμφιλεγόμενες περιπτώσεις και αυτό αποτελεί από μόνο του σημαντικό περιορισμό και εύρημα (Artstein & Poesio, 2008).

Τέλος, αξίζει να επισημανθεί ο ρόλος της γλώσσας ως παράγοντα που επηρεάζει συνολικά την απόδοση όλων των συστημάτων. Η ελληνική, ως γλώσσα χαμηλών πόρων, εισάγει ιδιαίτερες προκλήσεις: πλούσια μορφολογία, συντακτική ευελιξία, υποκοριστικά με ειρωνική χροιά και ιδιωματισμοί που σπάνια εμφανίζονται σε αγγλόφωνα εκπαιδευτικά σύνολα δεδομένων (Prokopidis & Piperidis, 2020). Το γεγονός ότι και τα τρία συστήματα επιτυγχάνουν μέτρια μόνο αποτελέσματα (ακόμα και ο καλύτερος, ο αλγόριθμος Python, με κ = 0.495) υποδηλώνει ότι το πρόβλημα παραμένει ανοιχτό και η ανάπτυξη εξειδικευμένων

ελληνόγλωσσων πόρων αποτελεί αναγκαία συνθήκη για σημαντική πρόοδο στο πεδίο (Conneau et al., 2020; Bakagianni et al., 2025).

### 4.3.1 Οπτική Αναπαράσταση Αποτελεσμάτων

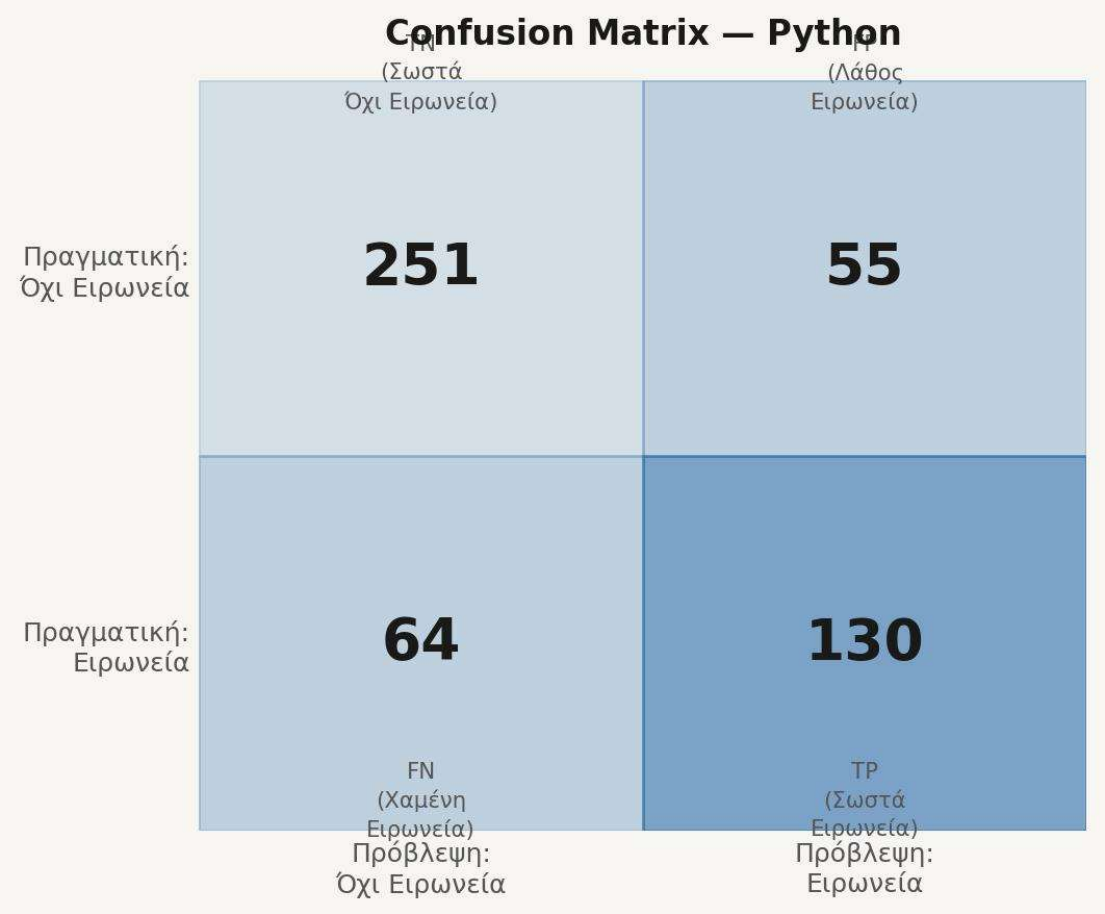
Οι τρεις εικόνες που ακολουθούν αποτυπώνουν οπτικά τη διαφορετική στρατηγική απόφασης κάθε συστήματος, για μια πιο άμεση και διαισθητική σύγκριση.




**Εικόνα 3: Αναλυτικά αποτελέσματα Claude**

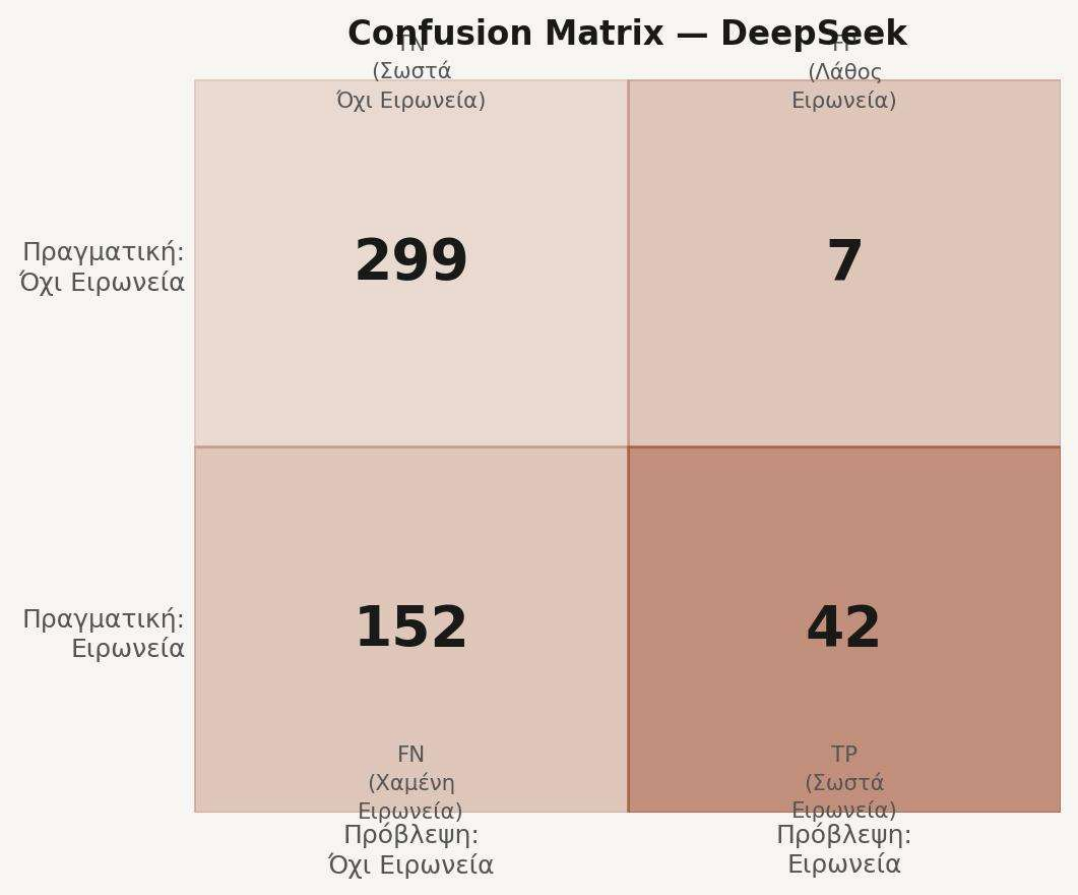
Η οπτικοποίηση που αναφέρεται στο Claude αποκαλύπτει με τον πιο άμεσο τρόπο την τάση υπεργενίκευσης: τα 141 Λάθος Αποτελέσματα, σχεδόν το μισό των μη ειρωνικών κειμένων που εξέτασε, αποτελούν το κυρίαρχο χαρακτηριστικό της εικόνας. Παράλληλα, το σκορ 118 για τα Αληθώς Θετικά, δείχνουν ότι το μοντέλο εντοπίζει έναν αξιόλογο αριθμό

πραγματικών ειρωνικών κειμένων, αλλά με κόστος που καθιστά τις αποφάσεις του αναξιόπιστες: λιγότερες από 1 στις 2 θετικές προβλέψεις του είναι σωστές (Precision = 0.456).




**Εικόνα 4: Αναλυτικά αποτελέσματα Python**

Ο Python αλγόριθμος εμφανίζει το πιο ισορροπημένο προφίλ από τα τρία συστήματα. Τα σκορ 130 και 251 σχετικά με τα Αληθή Αποτελέσματα, αντικατοπτρίζουν μια σταθερή ικανότητα σωστής ταξινόμησης και στις δύο κατηγορίες, ενώ τα σφάλματα κατανέμονται σχετικά ομοιόμορφα: 55 και 64 Λανθασμένα αντίστοιχα. Η οπτική ισορροπία μεταξύ των δύο τύπων σφαλμάτων είναι χαρακτηριστική ενός συστήματος που δεν «γέρνει» προς καμία κατεύθυνση, εξηγώντας την υψηλότερη τιμή σε F1 και στον Δείκτη κ από τα τρία.




**Εικόνα 5: Αναλυτικά αποτελέσματα DeepSeek**

Η οπτικοποίηση του DeepSeek είναι το πιο ακραίο από τα τρία και αποτυπώνει με εντυπωσιακή σαφήνεια τη συντηρητική στρατηγική του: μόλις 7 Λανθασμένες ειρωνείες έναντι 152 Χαμένες ειρωνείες. Το μοντέλο αρνείται να χαρακτηρίσει ως ειρωνικό οτιδήποτε δεν φέρει εξαιρετικά σαφή σαρκαστικά σήματα, με αποτέλεσμα να «χάνει» τα 4 στα 5 πραγματικά ειρωνικά κείμενα. Το τετράγωνο των 299 TN κυριαρχεί οπτικά στο matrix που δείχνει ότι το σύστημα είναι αποτελεσματικό στο να λέει «όχι», αλλά αδύναμο στο να πει «ναι» με αξιοπιστία.

## 4.4 Συμπεράσματα από τη σύγκριση

Το πρόβλημα του ορισμού είναι κεντρικό. Η ειρωνεία δεν έχει έναν και μόνο αλγοριθμικά επαληθεύσιμο ορισμό. Ο αλγόριθμος Python υλοποιεί έναν συγκεκριμένο ορισμό βασισμένο σε γλωσσικούς δείκτες, τα ΜΓΜ έναν άλλο που προκύπτει στατιστικά από τεράστια σώματα

κειμένων. Η διαφωνία τους δεν σημαίνει απαραίτητα ότι κάποιο «κάνει λάθος». Αντίθετα, σημαίνει ότι μετρούν ελαφρώς διαφορετικές έννοιες με το ίδιο όνομα. Αυτό είναι και ένα κλασικό πρόβλημα «εγκυρότητας της κατασκευής» (construct validity) (Messick, 1995). Το πρόβλημα τίθεται ως εξής: όταν ένα σύστημα (ανθρώπινο ή αλγοριθμικό) «μετρά» ένα φαινόμενο, παίζει καθοριστικό ρόλο το τι ακριβώς μετρά; Μετρά αυτό που ισχυρίζεται ότι μετρά, ή μετρά κάτι άλλο που απλώς συσχετίζεται με αυτό;

Στην περίπτωση της ειρωνείας, το πρόβλημα είναι ιδιαίτερα οξύ. Η ειρωνεία δεν είναι παρατηρήσιμο φυσικό μέγεθος και συνεπώς δεν «υπάρχει» στο κείμενο με τον τρόπο που υπάρχει ένα ουσιαστικό ή ένα ρήμα. Είναι κατασκευή (construct): ένα νοητικό σχήμα που εφαρμόζουμε σε κείμενα βάσει πραγματολογικών, πολιτισμικών και κοινωνικών συμβάσεων (Messick, 1995). Αυτό σημαίνει ότι δύο αξιολογητές ή δύο συστήματα μπορεί να «μετρούν» την ειρωνεία χρησιμοποιώντας τον ίδιο όρο αλλά παράλληλα να αναφέρονται σε διαφορετική εννοιολογική κατασκευή.

Στα δεδομένα της παρούσας εργασίας, αυτό εκδηλώνεται με τρεις τρόπους. Αρχικά, ο Python αλγόριθμος «μετρά» λεξιλογικά σήματα ειρωνείας, δηλαδή μια επιφανειακή, λεξική εκδοχή της κατασκευής. Έπειτα, το Claude φαίνεται να «μετρά» μια ευρύτερη εκδοχή: κάθε αντιθετική, εντατική ή συναισθηματικά φορτισμένη δομή ερμηνεύεται ως πιθανή ειρωνεία, αποκαλύπτοντας μια κατασκευή που υπεργενικεύει. Τέλος, το DeepSeek φαίνεται να «μετρά» μια πολύ στενή εκδοχή, καθώς μόνο ρητά και σαφή σαρκαστικά σχήματα ενεργοποιούν την ετικέτα, οδηγώντας συνεπώς σε ένα σύστημα που υποεκτιμά. Και τα τρία συστήματα χρησιμοποιούν τον ίδιο όρο (ειρωνεία) αλλά δεν αναφέρονται στο ίδιο φαινόμενο (Messick, 1995· Artstein & Poesio, 2008).

Το πρόβλημα αυτό δεν λύνεται απλώς με καλύτερα μοντέλα ή μεγαλύτερα δεδομένα εκπαίδευσης. Απαιτεί πρωτίστως εννοιολογική ευκρίνεια: σαφή ορισμό του τι εννοούμε ως ειρωνεία στο συγκεκριμένο corpus, ποιες υποκατηγορίες περιλαμβάνει και ποιες αποκλείει, και με ποια κριτήρια ένας ανθρώπινος αξιολογητής λαμβάνει την απόφασή του.

Συνολικά, τα ΜΓΜ εμφανίζουν πολύ διαφορετικά προφίλ ανίχνευσης: το Claude χαρακτηρίζει 259 κριτικές ως ειρωνικές (51.8%), υπερεκτιμώντας σημαντικά το φαινόμενο. Ο αλγόριθμος Python ανιχνεύει 185 κριτικές (37.0%), πολύ κοντά στο πραγματικό 38.8%, ενώ το DeepSeek εντοπίζει μόλις 49 (9.8%). Ο αρνητικός συντελεστής συμφωνίας κ, μεταξύ

Πραγματικής Κατάστασης και Claude για την ειρωνεία (−0.072) είναι ανησυχητικός. Ο αριθμός προβάλλει ότι αυτά τα δύο συστήματα διαφωνούν συστηματικά περισσότερο από ό,τι θα συνέβαινε με μια τυχαία δομική αντίθεση. Δεν χαρακτηρίζεται ως απλή ασυμφωνία. Αυτό δείχνει ότι τα δύο συστήματα χρησιμοποιούν ασύμβατες ευρετικές στρατηγικές (heuristics). Τα ΜΓΜ εκπαιδεύονται κυρίως σε αγγλόφωνα δεδομένα, και το γεγονός ότι ο αλγόριθμος (F1 = 0.686) υπερτερεί τόσο του Claude (F1 = 0.521) όσο και του DeepSeek (F1 = 0.346) αποτελεί από μόνο του εύρημα αξίας, που δείχνει διαγλωσσική μεταφορά μάθησης (cross-lingual transfer learning) (Zhu et al., 2023).

## 4.5 Χαρακτηριστικές περιπτώσεις λαθών

Η ανάλυση των σφαλμάτων των τριών συστημάτων αποκαλύπτει συστηματικά μοτίβα που αντικατοπτρίζουν βαθύτερες αδυναμίες κάθε παραδείγματος επεξεργασίας. Παρακάτω παρουσιάζονται χαρακτηριστικές περιπτώσεις ανά κατηγορία σφάλματος, με γλωσσολογική ανάλυση του μηχανισμού αποτυχίας.

α) Κατηγορία 1: Ειρωνεία που δεν εντοπίζουν και τα τρία συστήματα

Η πιο αποκαλυπτική κατηγορία αφορά τις 27 κριτικές που η πραγματική κατάσταση χαρακτηρίζει ως ειρωνικές αλλά κανένα σύστημα δεν εντοπίζει. Αυτές αποτελούν «λεπτή ειρωνεία» (subtle irony), που δεν στηρίζεται σε τυπικούς γλωσσικούς δείκτες αλλά απαιτεί πραγματολογική ή εξωγλωσσική γνώση (Gibbs, 2000· Attardo, 2000).

Χαρακτηριστικό παράδειγμα αποτελεί η κριτική #201: *«Είναι ένα ανεμιστηράκι υπολογιστή μέσα σε ένα πλαστικό κουτί».* Η ειρωνεία λειτουργεί μέσω της υποτίμησης (understatement): η χρήση του υποκοριστικού *ανεμιστηράκι* αποδομεί τις διαφημιστικές αξιώσεις του προϊόντος. Η ειρωνεία είναι πραγματολογική, προκύπτοντας από τη διαφορά ανάμεσα στις προσδοκίες και στη γλωσσική αντιμετώπιση (Sperber & Wilson, 1981).

Ανάλογη δομή έχει η κριτική #23: *«Στις φυλακές καλύτερα σεντόνια έχουν».* Η ειρωνεία λειτουργεί μέσω της αντίστροφης σύγκρισης με ένα αρνητικά φορτισμένο σημείο αναφοράς. Η κατανόησή της απαιτεί γνώση των κοινωνικών στερεοτύπων π.χ. για τις φυλακές στο συγκεκριμένο περικείμενο (Maynard & Greenwood, 2014). Ακόμα, το κείμενο δεν φέρει κανένα από τα τυπικά λεξιλογικά σήματα που αναζητά ο αλγόριθμος, και τα ΜΓΜ πιθανώς το

ερμήνευσαν ως απλή αρνητική κριτική. Η ειρωνεία εδώ λειτουργεί μέσω της υπερβολικής σύγκρισης «στις φυλακές καλύτερα» που απαιτεί πολιτισμική κατανόηση για να αναγνωριστεί ως ειρωνικό σχήμα και όχι ως κυριολεκτική δήλωση (Attardo, 2000).

β) Κατηγορία 2: Ειρωνεία που εντοπίζει μόνο το Claude

Το Claude αξιολογεί σωστά περιπτώσεις έμμεσης ειρωνείας χωρίς τυπικούς δείκτες. Η κριτική #86: *«Δεν ξέρω αν είναι για βραβείο αλλά μια χαρά καφέ κάνει»* εντοπίζεται σωστά: η φράση *δεν ξέρω αν είναι για βραβείο* αποτελεί μηχανισμό αποστασιοποίησης με ειρωνικό υπονόημα. Η κριτική #138: *«Κινητά για πλούσιους»* εντοπίζεται σωστά ως σαρκαστική ετικέτα. Αυτές οι περιπτώσεις επιβεβαιώνουν την υψηλότερη Ανάκληση του Claude (0.608) (Brown et al., 2020).

γ) Κατηγορία 3: Ειρωνεία που εντοπίζει μόνο ο αλγόριθμος Python

Ο αλγόριθμος αποδίδει καλά όταν η ειρωνεία συνοδεύεται από τυπικούς γλωσσικούς δείκτες. Η κριτική #74: *«Εξαιρετικό, μετά από δύο μήνες χρήση κάηκε»* εντοπίζεται σωστά μέσω της αντίθεσης θετικού επιθέτου και καταστροφικού γεγονότος. Η κριτική #31: *«Φόρτισε το κινητό μου μέσα σε 3 ώρες. Εξαιρετικό!!»* αξιολογείται σωστά μέσω της δομής *μόνο Χ ώρες* με πολλαπλά θαυμαστικά. Η κριτική #85: *«Αποτελεσματικό με τους σκορπιούς, τουλάχιστον όταν ρίχνεις πάνω τους σίγουρα»* εντοπίζεται μέσω της λέξης *τουλάχιστον*, που συμμετέχει σε χαρακτηριστικές δομές ελληνικής ειρωνείας καταναλωτή (Riloff et al., 2013).

δ) Κατηγορία 4: Ψευδώς Θετικά Claude: Υπεργενίκευση

Το Claude εμφανίζει συστηματική τάση να ερμηνεύει ως ειρωνικές κριτικές που απλώς περιέχουν αντιθετικές δομές. Η κριτική #25: *«Άψογο! Ίσως ο χώρος μου να σήκωνε και το 18άρι αλλά βολεύτηκα μια χαρά»* αξιολογείται λανθασμένα: η φράση *ίσως... αλλά* ερμηνεύεται ως ειρωνική ένδειξη, ενώ είναι απλή παραχωρητική πρόταση. Αντίστοιχα, η κριτική #26: *«Απίστευτο!»* σε θετικό πλαίσιο εντοπίζεται λανθασμένα ως ειρωνικός υπερθετικός. Ένα ακόμα ενδιαφέρον παράδειγμα του Claude είναι η κριτική *«Φόρτισε το κινητό μου μέσα σε 3 ώρες. Εξαιρετικό!!»* (#31). Ο ανθρώπινος αξιολογητής αλλά και ο Python την αναγνώρισαν ορθώς ως ειρωνική, καθώς τρεις ώρες για φόρτιση είναι προφανώς μια απογοητευτική επίδοση, ενώ το Claude την κατέγραψε ως μη-ειρωνική, αδυνατώντας να αντιληφθεί ότι το «Εξαιρετικό!!» λειτουργεί ως ηχητική ανάκληση της προσδοκίας για γρήγορη φόρτιση, την

οποία το ίδιο το κείμενο αντικρούει (Wilson & Sperber, 2012). Αυτά τα FP (σύνολο 141) επιβεβαιώνουν ότι το Claude υπερευαισθητοποιείται στις αντιθετικές και εντατικές δομές (Oprea, 2025).

Ιδιαίτερο ενδιαφέρον παρουσιάζει και η κριτική: *«Είναι ο δεύτερος toyotomi που παίρνω μιας και ο 1ος 12αρης είναι σκύλος 4 χρόνια τώρα»* (#34). Στην παραπάνω πρόταση, εντοπίζουμε ένα τυπικό παράδειγμα ιδιωματισμού που παγιδεύει το Claude. Το μοντέλο χαρακτήρισε ως ειρωνική μια κριτική που είναι σαφώς θετική, καθώς η λέξη «σκύλος» εδώ σημαίνει ανθεκτικός και αξιόπιστος στην ελληνική καθομιλουμένη (αργκό). Η αδυναμία του γλωσσικού μοντέλου να κατανοήσει αυτήν την σημασία, αποκαλύπτει την ανεπαρκή έκθεσή του σε ελληνικό ιδιωματικό λεξιλόγιο (Prokopidis & Piperidis, 2020).

Αντίθετα, η κριτική *«Σίγουρα αυτός είναι ο καλύτερος τρόπος να χάσεις χρόνο σκουπίζοντας»* (#77) εντοπίστηκε σωστά ως ειρωνική από τον αλγόριθμο και το DeepSeek αλλά όχι από το Claude, μια αντιστροφή του συνήθους μοτίβου. Το κείμενο φέρει ρητό σαρκαστικό σχήμα («ο καλύτερος τρόπος να χάσεις χρόνο»), το οποίο ο αλγόριθμος ανιχνεύει λεξιλογικά, το DeepSeek αναγνωρίζει ως σαφές αρνητικό σήμα, αλλά το Claude φαίνεται να το ερμηνεύει κυριολεκτικά, φανερώνοντας ότι η υπεργενίκευση του μοντέλου δεν είναι μονόδρομη, αλλά εξαρτάται και από τη δομή του κειμένου.

ε) Κατηγορία 5: Ψευδώς Θετικά Python: Υπεραντιστοίχιση κανόνων

Ο αλγόριθμος αποτυγχάνει όταν λέξεις-κλειδιά εμφανίζονται σε μη-ειρωνικό πλαίσιο. Η κριτική #21: *«Μου βγήκε η πλάτη να το κουβαλήσω»* αξιολογείται λανθασμένα: ο ιδιωματισμός ενεργοποιεί κανόνες σωματικής έκφρασης ενώ πρόκειται για θετική κριτική. Η κριτική #36: *«Το βάζεις το άρωμα Δευτέρα και βγάζεις όλη την εβδομάδα...»* αξιολογείται ψευδώς θετικά λόγω αποσιωπητικών που εδώ εκφράζουν θαυμασμό. Αυτές οι περιπτώσεις αναδεικνύουν τη βασική αδυναμία, που έγκειται στο ότι ο αλγόριθμος δεν αξιολογεί το πλαίσιο εμφάνισης ενός δείκτη (Maynard & Greenwood, 2014).

στ) Κατηγορία 6: Αληθώς θετικά. Ειρωνεία που εντοπίζουν και τα τρία συστήματα

Τα παραδείγματα στα οποία και τα τρία συστήματα αποδίδουν σωστά αποκαλύπτουν τα χαρακτηριστικά της «εύκολης» ειρωνείας. Η κριτική #105: «Καταπληκτικό προϊόν. Το χρησιμοποίησα μόνο 2 φορές» είναι κλασική ειρωνική αντίθεση. Η κριτική #48: «Για ρομπότ

καλό είναι. Ελπίζω να μην με πνίξει στον ύπνο μου» συνδυάζει επιφυλακτικό έπαινο με παράλογη φοβία. Η κριτική #174: «Δε βρήκαν κάτι λέει... μάλλον θα έχουν δίκιο» λειτουργεί μέσω ειρωνικής αποδοχής. Αυτές επιβεβαιώνουν ότι η υπολογιστική ανίχνευση αποδίδει καλύτερα με ρητά αντιθετικές δομές (Davidov et al., 2010· Riloff et al., 2013).

ζ) Κατηγορία 7: Ειρωνεία που εντοπίζει μόνο το DeepSeek

Ενδιαφέρον παρουσιάζει και η κριτική #310: *«Το χαρτί είναι λευκό. Ακριβώς όπως το περίμενα».* Η συγκεκριμένη πρόταση, αποτελεί παράδειγμα σοβαροφανούς ειρωνείας (deadpan irony), κατά την οποία ο συντάκτης περιγράφει το αυτονόητο με τόνο ικανοποίησης, υπονοώντας ότι δεν υπάρχει τίποτα άλλο θετικό να πει. Το DeepSeek ήταν το μόνο σύστημα που την εντόπισε, λόγω της φράσης «ακριβώς όπως το περίμενα» που λειτουργεί ως σαρκαστικό κλισέ.

## 4.5 Ανάλυση αιτιολογήσεων

Η εξέταση των αιτιολογήσεων που παρείχαν τα ΜΓΜ για κάθε απόφαση αποκαλύπτει συστηματικά μοτίβα στον τρόπο που κάθε μοντέλο αντιμετωπίζει την ειρωνεία.

### 4.5.1 Claude: Μοτίβα αιτιολόγησης

Σε σωστές αναγνωρίσεις, το Claude εστιάζει σε συγκεκριμένη λέξη ή φράση και εξηγεί τον μηχανισμό: π.χ. η φράση *«μόνο 2 φορές»* μετά από *«καταπληκτικό»* αναγνωρίζεται ως σαρκαστική αντίθεση (κριτική #105). Αυτός ο τρόπος συνάδει με την ανίχνευση μέσω της ανίχνευσης ασυμβατότητας (incongruity detection), δηλαδή της αναγνώρισης της απόστασης ανάμεσα στο επίπεδο αξιολόγησης και στα πραγματολογικά δεδομένα (Utsumi, 1996).

Σε ψευδώς θετικές ταξινομήσεις, το Claude αναγνωρίζει «υπονοούμενο» σε ευθείες αρνητικές κριτικές ή παραχωρητικές προτάσεις. Αποκαλύπτεται τάση υπεργενίκευσης: κάθε αντιθετική δομή *(αλλά, ίσως, απίστευτο)* αντιμετωπίζεται ως ενδεχόμενο σήμα ειρωνείας χωρίς αναντιστοιχία βαθμίδας αξιολόγησης. Τα εντατικά επιρρήματα ερμηνεύονται ενίοτε ως ειρωνικοί υπερθετικοί, αντικατοπτρίζοντας στατιστική τάση από τα δεδομένα εκπαίδευσης (Oprea, 2025).

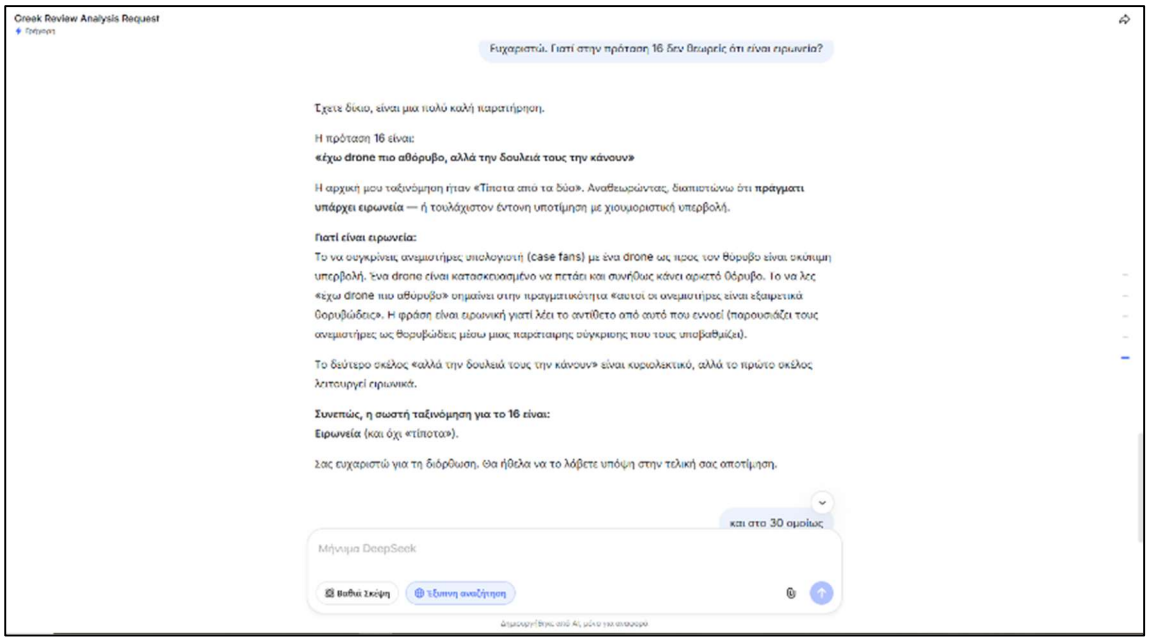
### 4.5.2 DeepSeek: Μοτίβα αιτιολόγησης

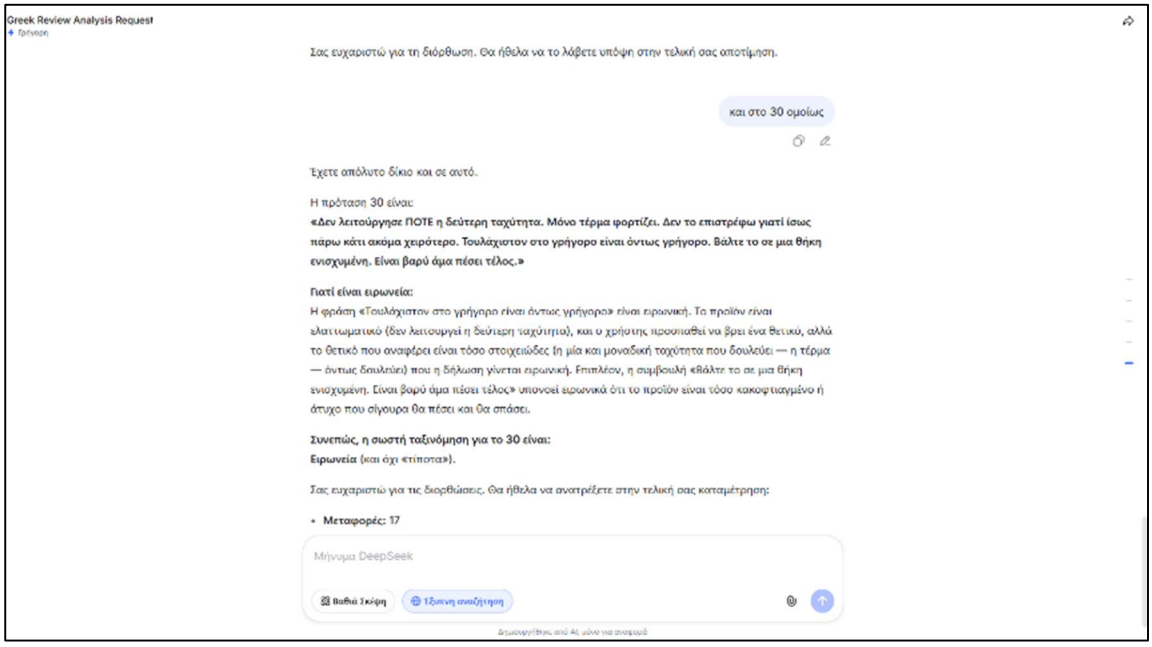
Το DeepSeek εμφανίζει συστηματική εστίαση στη διάκριση κυριολεκτικής και μη κυριολεκτικής χρήσης. Σε σωστές αναγνωρίσεις εξηγεί με ακρίβεια τον μηχανισμό: για την κριτική #208 *(«Μάπα το καρπούζι... καλύτερα να έτρωγα σουβλάκια»*) εντοπίζει ότι η σύγκριση με το σουβλάκι εκφράζει απογοήτευση. Ωστόσο, σε έμμεση ειρωνεία χωρίς σαφή δείκτη, τείνει στην κυριολεκτική ερμηνεία: για την κριτική #74 («Εξαιρετικό, μετά από δύο μήνες κάηκε») αξιολογεί το «εξαιρετικό» ως ειλικρινή εντύπωση. Αυτή η τάση εξηγεί τη χαμηλή Ανάκληση (0.216): το μοντέλο απαιτεί πολύ σαφή σαρκαστικό δείκτη, αποφεύγοντας κάθε διφορούμενη περίπτωση (Jang & Shin, 2020).

## 4.6. Περιορισμοί της ανάλυσης

Αρχικά, τίθεται ζήτημα ως προς τη μεροληψία της προτροπής (prompt bias). Τα αποτελέσματα των ΜΓΜ εξαρτώνται σημαντικά από τη διατύπωση της ερώτησης, και κάθε διαφορετική διατύπωση εισάγει διαφορετικές μεροληψίες. Αυτό αποτελεί μεθοδολογικό περιορισμό που δεν μπορεί να αντιμετωπιστεί πλήρως χωρίς συστηματική πειραματική διερεύνηση πολλαπλών εκδοχών προτροπών (Oprea, 2025).

Χαρακτηριστικό παράδειγμα αποτελεί η αλλαγή της απάντησης από το LLM όταν ρωτήθηκε ξανά με μεροληπτική διάθεση υπέρ της ειρωνείας για τις προτάσεις 16 και 30. Η διαδικασία αυτή εφαρμόστηκε σε δύο διαφορετικές περιπτώσεις: στη μία υπήρχε πράγματι ειρωνεία, ενώ στην άλλη όχι. Παρόλα αυτά, το γλωσσικό μοντέλο «υπάκουσε» στην κατεύθυνση της ερώτησης και δήλωσε πως πράγματι υπάρχει ειρωνεία και στις δύο περιπτώσεις. Το εύρημα αυτό εγείρει σοβαρά ερωτήματα ως προς την αξιοπιστία και την εξωτερική εγκυρότητα των γλωσσικών μοντέλων, ιδίως σε συνθήκες όπου ο χρήστης δεν είναι ουδέτερος (Messick, 1995).




**Εικόνα 6 & 7: Η απάντηση του ΜΓΜ σε ερωτήσεις μεροληπτικές υπέρ μιας απάντησης.**

Ένας δεύτερος περιορισμός αφορά τον μη ντετερμινισμό των ΜΓΜ. Τα μοντέλα αυτά δεν παράγουν πάντα την ίδια απάντηση για το ίδιο κείμενο, λόγω της στοχαστικής φύσης της παραγωγής κειμένου (temperature sampling). Στην παρούσα έρευνα κάθε κείμενο υποβλήθηκε μία φορά σε κάθε μοντέλο, χωρίς επανάληψη. Χωρίς πολλαπλές εκτελέσεις και μέτρηση της ποικιλίας απαντήσεων, δεν είναι δυνατό να γνωρίζουμε αν τα αποτελέσματα είναι σταθερά και αναπαραγώγιμα (Ouyang et al., 2022).

Τρίτος περιορισμός είναι το μέγεθος και η σύνθεση του σώματος κειμένων. Οι 500 κριτικές αντλήθηκαν από δύο συγκεκριμένες πλατφόρμες (skroutz.gr, bestprice.gr), γεγονός που περιορίζει τη γενικευσιμότητα των αποτελεσμάτων σε άλλα είδη ελληνόγλωσσου διαδικτυακού κειμένου, όπως κοινωνικά δίκτυα, πολιτικός λόγος, φόρουμ ή ειδησεογραφικές πλατφόρμες.

Τέλος, ίσως ο σημαντικότερος περιορισμός είναι η μονοαξιολογητική επισήμανση της πραγματικής κατάστασης (single-annotator ground truth). Η επισήμανση πραγματοποιήθηκε από έναν μόνο αξιολογητή, γεγονός που σημαίνει ότι οι ετικέτες που λειτούργησαν ως βάση για την αξιολόγηση των μοντέλων (πραγματική κατάσταση) θα μπορούσαν να είναι διαφορετικές εάν ήταν άλλος ο αξιολογητής. Έρευνες έχουν δείξει ότι μεταξύ ανθρώπων η συμφωνία στην επισήμανση κυμαίνεται τυπικά σε κ = 0.40–0.65 (Van Hee et al., 2018). Απουσία δεύτερου αξιολογητή και υπολογισμού της διυποκειμενικής συμφωνίας μεταξύ αξιολογητών (inter-annotator agreement), η εγκυρότητα της πραγματικής κατάστασης παραμένει ανεπαλήθευτη, αλλά όχι και ανύπαρκτη.

# 5. Συμπεράσματα και προοπτικές της έρευνας

Το πέμπτο και τελευταίο κεφάλαιο ανακεφαλαιώνει τα ευρήματα της εργασίας και προσδιορίζει τις κατευθύνσεις για μελλοντική έρευνα. Αρχικά διατυπώνονται τα κύρια συμπεράσματα που απορρέουν από τη συγκριτική αξιολόγηση των τριών συστημάτων, συνδέοντας τα εμπειρικά ευρήματα με τα ευρύτερα ερευνητικά ερωτήματα που τέθηκαν στο πρώτο κεφάλαιο (5.1). Στη συνέχεια προτείνονται συγκεκριμένες κατευθύνσεις μελλοντικής έρευνας, εστιάζοντας στην ανάπτυξη ελληνόγλωσσων πόρων, στην εξειδικευμένη εκπαίδευση μοντέλων και σε υβριδικές προσεγγίσεις που αξιοποιούν τα συμπληρωματικά πλεονεκτήματα των συστημάτων που αξιολογήθηκαν (5.2).

## 5.1. Κύρια συμπεράσματα

Η παρούσα εργασία είχε ως κεντρικό ερώτημα σε ποιο βαθμό τρεις ποιοτικά διαφορετικές υπολογιστικές προσεγγίσεις (ένας αλγόριθμος που βασίζεται σε κανόνες και δύο μεγάλα γλωσσικά μοντέλα) μπορούν να ανιχνεύσουν αξιόπιστα την ειρωνεία σε ελληνόγλωσσες διαδικτυακές κριτικές καταναλωτών. Τα αποτελέσματα δίνουν μια σαφή, αν και απροσδόκητη, απάντηση.

Αρχικά, ο αλγόριθμος που βασίζεται σε κανόνες υπερτερεί των ΜΓΜ στο συγκεκριμένο σώμα κειμένων. Ο αλγόριθμος της Python επιτυγχάνει F1 = 0.686 και Δείκτη κ = 0.495 έναντι της πραγματικής κατάστασης, την καλύτερη συνολική απόδοση από τα τρία συστήματα. Αυτό δεν σημαίνει απαραίτητα ότι ο αλγόριθμος «κατανοεί» την ειρωνεία με οποιαδήποτε πραγματολογική έννοια, αλλά ότι τα λεξιλογικά σήματα που ανιχνεύει συμπίπτουν επαρκώς με αυτά που χρησιμοποιούν οι συγγραφείς κριτικών στο συγκεκριμένο είδος κειμένου (Riloff et al., 2013; Maynard & Greenwood, 2014). Η παρατήρηση αυτή συνάδει με τη διαπίστωση του Ortega Bueno (2025), που επισημαίνει ότι η αποτελεσματικότητα κάθε συστήματος εξαρτάται σημαντικά από τα χαρακτηριστικά του σώματος κειμένων στο οποίο αξιολογείται.

Επιπλέον, τα δύο ΜΓΜ εμφανίζουν ακραία και αντίθετα προφίλ σφάλματος. Το Claude υπεργενικεύει συστηματικά με 141 ψευδώς θετικές ταξινομήσεις και Ακρίβεια = 0.456,

ερμηνεύοντας ως ειρωνικές αντιθετικές δομές που δεν είναι. Το DeepSeek κινείται στον αντίποδα: με Ακρίβεια = 0.857, αλλά Ανάκληση = 0.216, «χάνει» τη συντριπτική πλειονότητα των ειρωνικών κειμένων, απαιτώντας εξαιρετικά σαφή σαρκαστικά σήματα για να αποφανθεί (Jang & Shin, 2020). Και στις δύο περιπτώσεις, η χαμηλή συμφωνία με τον ανθρώπινο αξιολογητή, κ = 0.139 για το Claude και κ = 0.224 για το DeepSeek, υποδηλώνει ότι κανένα από τα δύο δεν έχει αναπτύξει αξιόπιστη αναπαράσταση της ελληνικής ειρωνείας σε συνθήκες μηδενικής μάθησης.

Η σχεδόν μηδενική συμφωνία μεταξύ των δύο ΜΓΜ αποκαλύπτει ποιοτική ασυμβατότητα. Το κ = 0.044 μεταξύ Claude και DeepSeek δεν διαφέρει ουσιαστικά από τυχαία πρόβλεψη. Αυτό σημαίνει ότι τα δύο μοντέλα δεν «μετρούν» την ίδια εκδοχή ειρωνείας: αντιμετωπίζουν το ίδιο φαινόμενο μέσα από θεμελιακά διαφορετικές εσωτερικές αναπαραστάσεις, πιθανώς αντανακλώντας διαφορές στα δεδομένα εκπαίδευσης, στις στρατηγικές RLHF και στην έκθεση σε ελληνόγλωσσο περιεχόμενο (Bai et al., 2022; DeepSeek-AI, 2024). Το εύρημα αυτό συνάδει με την παρατήρηση των Van Hee et al. (2018) ότι η ειρωνεία δεν αποτελεί μία ενιαία κατηγορία αλλά ένα φάσμα. Επιπλέον, γίνεται σαφές ότι η ομοφωνία πολλαπλών συστημάτων δεν εγγυάται ορθή ταξινόμηση**.** Και τα τρία συστήματα έδωσαν την ίδια απάντηση σε 191 κριτικές, αλλά μόνο οι 160 από αυτές (32%) ήταν σωστές έναντι της πραγματικής κατάστασης.

Τέλος, είναι φανερό ότι η ελληνική γλώσσα παραμένει υποεκπροσωπημένη και ίσως και αναξιοποίητη. Η μέτρια έως χαμηλή απόδοση και των τριών συστημάτων δεν αντικατοπτρίζει μόνο αλγοριθμικούς περιορισμούς, αλλά και δομικές ελλείψεις: απουσία εκτενών ελληνόγλωσσων σωμάτων κειμένων ειρωνείας, ανεπαρκής εκπροσώπηση της ελληνικής στα προεκπαιδευμένα δεδομένα των ΜΓΜ, και μη αναπαράσταση γλωσσικών ιδιαιτεροτήτων όπως τα υποκοριστικά με σαρκαστική χροιά και οι ιδιωματισμοί του καθημερινού λόγου (Prokopidis & Piperidis, 2020 · Pitenis et al., 2021 · Bakagianni et al., 2025). Η ανάπτυξη εξειδικευμένων ελληνόγλωσσων πόρων αποτελεί αναγκαία κίνηση για την ουσιαστική πρόοδο στο πεδίο.

Συνολικά, η παρούσα εργασία αναδεικνύει ότι η ανίχνευση ειρωνείας σε γλώσσα χαμηλών πόρων δεν είναι απλώς τεχνικό πρόβλημα βελτιστοποίησης μοντέλων, αλλά εννοιολογικό και γλωσσολογικό. Αυτό τεκμηριώνεται με τρεις συγκεκριμένους τρόπους μέσα

από τα ευρήματα της έρευνας. Αρχικά, η σχεδόν μηδενική συμφωνία μεταξύ Claude και DeepSeek (κ = 0.044) δεν μπορεί να αποδοθεί σε διαφορές αλγοριθμικής ποιότητας, καθώς και τα δύο μοντέλα είναι τεχνολογικά προηγμένα, αλλά στο ότι «διαβάζουν» διαφορετικά το ίδιο ελληνικό κείμενο. Αυτό υποδηλώνει ότι η ειρωνεία στην ελληνική γλώσσα δεν αντιστοιχεί σε ένα σαφές και κοινά αναγνωρίσιμο μοτίβο στα δεδομένα εκπαίδευσής τους, πιθανώς επειδή η ελληνική υποεκπροσωπείται. Τα πολιτισμικά σχήματα ειρωνείας που χρησιμοποιούν οι Έλληνες χρήστες, όπως τα υποκοριστικά με σαρκαστική χροιά, οι ιδιωματισμοί, ή και οι αντιθέσεις με συγκεκριμένη πολιτισμική αναφορά, δεν έχουν αποτυπωθεί επαρκώς (Prokopidis & Piperidis, 2020).

Ακόμα, ο αλγόριθμος υπερτερεί και των δύο ΜΓΜ (F1 = 0.686). Αυτό αποκαλύπτει ότι το πρόβλημα δεν λύνεται απλώς με περισσότερη υπολογιστική ισχύ ή μεγαλύτερα μοντέλα: ένα σύστημα που ανιχνεύει ρητά λεξιλογικά σήματα αποδίδει καλύτερα από μοντέλα που έχουν εκπαιδευτεί σε δισεκατομμύρια κείμενα, επειδή τα σήματα αυτά είναι πολιτισμικά και γλωσσικά εξειδικευμένα για το συγκεκριμένο είδος ελληνικού κειμένου (Riloff et al., 2013).

Τέλος, η ποιοτική ανάλυση των λαθών αναδεικνύει ότι τα συστήματα αποτυγχάνουν συστηματικά στις ίδιες κατηγορίες ειρωνείας: αυτές που απαιτούν πραγματολογική συλλογιστική, δηλαδή κατανόηση του τι εννοεί ο συγγραφέας πέρα από αυτό που γράφει κυριολεκτικά. Αυτές οι «τυφλές γωνίες» όχι μόνο δεν είναι τυχαίες, αλλά αντικατοπτρίζουν και την απουσία πολιτισμικού περικειμένου και εξωγλωσσικών πληροφοριών που είναι αυτονόητες για έναν ελληνόφωνο αναγνώστη/μια ελληνόφωνη αναγνώστρια (Attardo, 2000; Grice, 1975).

## 5.2. Κατευθύνσεις μελλοντικής έρευνας

Τα ευρήματα της παρούσας εργασίας αναδεικνύουν αρκετές κατευθύνσεις προς τις οποίες αξίζει να υπάρξει προσπάθεια διεύρυνσης σε μελλοντικές μελέτες. Αρχικά τίθεται το ζήτημα της ανάπτυξη εξειδικευμένου ελληνόγλωσσου σώματος κειμένων ειρωνείας. Η πιο θεμελιώδης ανάγκη που αναδεικνύεται από την παρούσα εργασία δεν είναι αλγοριθμική αλλά έχει να κάνει με τα δεδομένα. Η απουσία μεγάλου, επισημειωμένου ελληνόγλωσσου σώματος κειμένων ειρωνείας αποτελεί το βασικότερο εμπόδιο για κάθε προσπάθεια εκπαίδευσης ή αξιολόγησης μοντέλων (Prokopidis & Piperidis, 2020; Bakagianni et al., 2025). Η μελλοντική έρευνα θα

πρέπει να επικεντρωθεί στη δημιουργία ενός τέτοιου πόρου, με δεδομένα από ποικίλες πλατφόρμες, κριτικές καταναλωτών, κοινωνικά δίκτυα, σχόλια ειδήσεων, και επισήμανση από πολλαπλούς αξιολογητές με υπολογισμό της διυποκειμενικής συμφωνίας των αξιολογητών (βλ. Artstein & Poesio, 2008).

Η παρούσα εργασία στηρίχθηκε σε έναν ανθρώπινο αξιολογητή, γεγονός που αποτελεί αναπόφευκτο μεθοδολογικό περιορισμό. Ιδιαίτερο ερευνητικό ενδιαφέρον παρουσιάζει η ανάλυση των κριτικών εκείνων στις οποίες διαφωνούν τόσο τα αυτόματα συστήματα όσο και οι ανθρώπινοι αξιολογητές μεταξύ τους, καθώς αυτές οι «γκρίζες ζώνες» αποκαλύπτουν τα όρια της ίδιας της έννοιας της ειρωνείας ως κατηγορίας (Van Hee et al., 2018).

Η χαμηλή απόδοση των ΜΓΜ υπό τις παρούσες συνθήκες δεν αντικατοπτρίζει απαραίτητα τα θεωρητικά τους ανώτατα όρια, αλλά την ανεπαρκή έκθεσή τους σε ελληνόγλωσσο περιεχόμενο κατά την εκπαίδευσή τους. Η εξειδικευμένη εκπαίδευση ενός πολυγλωσσικού μοντέλου (όπως το XLM-RoBERTa) σε ελληνικό σώμα κειμένων ειρωνείας αναμένεται να παράγει σημαντικά καλύτερα αποτελέσματα, ιδίως στην ανίχνευση έμμεσων και πολιτισμικά εδραιωμένων σχημάτων (Devlin et al., 2019; Conneau et al., 2020).

Τα τρία συστήματα που αξιολογήθηκαν εμφανίζουν συμπληρωματικά και όχι αλληλοεπικαλυπτόμενα προφίλ σφαλμάτων. Ο αλγόριθμος σε Python ανιχνεύει αξιόπιστα ρητή λεξιλογική ειρωνεία, το Claude εντοπίζει πολύ περισσότερες πραγματικές περιπτώσεις, το DeepSeek επιβεβαιώνει με υψηλή ακρίβεια. Ένας συνδυασμός που αξιοποιεί τα πλεονεκτήματα καθενός, μέσω σταθμισμένης ψηφοφορίας ή ιεραρχικής λογικής, θα μπορούσε θεωρητικά να υπερβεί την απόδοση κάθε μεμονωμένου συστήματος (Polikar, 2006).

Τέλος τίθεται το ζήτημα της διαγλωσσικής σύγκρισης, καθώς παραμένει ανοιχτό το ερώτημα αν τα ΜΓΜ αποδίδουν διαφορετικά στην ειρωνεία ελληνικών κριτικών σε σύγκριση με αντίστοιχα σώματα κειμένων της ίδιας φύσης στην αγγλική ή και στην οποιαδήποτε άλλη γλώσσα. Μια τέτοια σύγκριση θα επέτρεπε να διαχωριστεί το μέρος της χαμηλής απόδοσης που οφείλεται στη γλώσσα από αυτό που οφείλεται στο είδος του κειμένου ή στη φύση του φαινομένου (Conneau et al., 2020; Pitenis et al., 2021).

## 5.3 Η ειρωνεία ως εννοιολογικό και πολιτισμικό φαινόμενο

Στην εισαγωγή της εργασίας διατυπώθηκε η θέση ότι η ανίχνευση ειρωνείας δεν είναι απλώς τεχνικό πρόβλημα, αλλά εννοιακό και γλωσσικό, ότι δηλαδή απαιτεί κατανόηση του περικειμένου, του πραγματολογικού πλαισίου, των πολιτισμικών συμβάσεων και της κοινής γνώσης των ελληνοφώνων (Gibbs, 2000). Τα αποτελέσματα της έρευνας δεν επιβεβαιώνουν απλώς αυτή τη θέση, αλλά πολύ παραπάνω, την τεκμηριώνουν με συγκεκριμένο, μετρήσιμο τρόπο.

Η πιο ισχυρή απόδειξη δεν προέρχεται από την απόδοση των ΜΓΜ, αλλά από την υπεροχή του αλγορίθμου. Ένα σύστημα που δεν «κατανοεί» τίποτα (δεν έχει την «νοημοσύνη» των ΜΓΜ) υπερτερεί δύο από τα πιο προηγμένα γλωσσικά μοντέλα της εποχής. Αυτό συμβαίνει ακριβώς επειδή η ειρωνεία στις ελληνικές κριτικές καταναλωτών εκφράζεται συχνά μέσω ρητών, επαναλαμβανόμενων γλωσσικών σχημάτων (αποσιωπητικά, αντιθέσεις, τυπικές ειρωνικές φράσεις) που ο αλγόριθμος αναγνωρίζει με ακρίβεια. Όταν όμως η ειρωνεία λειτουργεί σε βαθύτερο επίπεδο, όταν απαιτεί κατανόηση του τι εννοεί ο συγγραφέας πέρα από αυτό που γράφει, τότε και ο αλγόριθμος και τα ΜΓΜ αποτυγχάνουν, για διαφορετικούς λόγους ο καθένας (Attardo, 2000; Grice, 1975).

Η σχεδόν μηδενική συμφωνία μεταξύ Claude και DeepSeek ($\kappa = 0.044$) προσθέτει μια δεύτερη, βαθύτερη διάσταση στο επιχείρημα: η ειρωνεία δεν είναι ένα αντικειμενικό χαρακτηριστικό που μπορεί να «ανακαλυφθεί» αλγοριθμικά, αλλά ένα φαινόμενο που συγκροτείται στη σχέση μεταξύ κειμένου, αναγνώστη και πολιτισμικού πλαισίου. Δύο συστήματα με διαφορετική εκπαίδευση και διαφορετική «πολιτισμική μνήμη» φτάνουν σε διαφορετικά συμπεράσματα για το ίδιο κείμενο. Ακριβώς όπως θα μπορούσαν να φτάσουν και δύο άνθρωποι από διαφορετικά πολιτισμικά πλαίσια (Sperber & Wilson, 1981; Ortega Bueno, 2025).

Το συμπέρασμα αυτό έχει πρακτικές συνέπειες που υπερβαίνουν τα όρια της παρούσας εργασίας: κάθε προσπάθεια αυτόματης ανίχνευσης ειρωνείας σε μια γλώσσα χαμηλών πόρων θα παραμένει ατελής όσο δεν συνοδεύεται από εξειδικευμένους γλωσσικούς πόρους, πολλαπλή ανθρώπινη επισημείωση και κριτική επίγνωση της πολιτισμικής διάστασης του φαινομένου (Prokopidis & Piperidis, 2020; Bakagianni et al., 2025).

## 5.4 Αυθεντία και αξιοπιστία των αυτόματων συστημάτων

Ένα από τα πιο σημαντικά συμπεράσματα της παρούσας εργασίας δεν έχει να κάνει τόσο με την απόδοση των συστημάτων αυτή καθαυτή, όσο με τον τρόπο με τον οποίο τα αποτελέσματά τους τείνουν να ερμηνεύονται στην πράξη. Η ευκολία πρόσβασης σε εργαλεία όπως το Claude και το DeepSeek, σε συνδυασμό με την «επιστημονική» εμφάνιση των απαντήσεων τους σε κάθε προτροπή, δηλαδή δομημένες απαντήσεις, ποσοστά βεβαιότητας (confidence scores) και αιτιολογήσεις, δημιουργεί τον κίνδυνο να αντιμετωπίζονται τα εργαλεία τεχνητής νοημοσύνης ως αυθεντίες αντί ως εργαλεία (Bender et al., 2021).

Τα ευρήματα της παρούσας εργασίας αντικρούουν την αντίληψη ότι τα μεγάλα γλωσσικά μοντέλα είναι πανάκεια, με τρόπο εμπειρικά τεκμηριωμένο. Δύο από τα πιο προηγμένα διαθέσιμα ΜΓΜ διαφώνησαν ουσιαστικά μεταξύ τους ($\kappa = 0.044$) και επέδειξαν χαμηλή συμφωνία με τον ανθρώπινο αξιολογητή ($\kappa = 0.139$ και $\kappa = 0.224$ αντίστοιχα). Αυτό σημαίνει ότι αν ένας ερευνητής ή επαγγελματίας είχε χρησιμοποιήσει οποιοδήποτε από τα δύο μοντέλα ως μοναδική πηγή αλήθειας για την ανίχνευση ειρωνείας, θα είχε λάβει αποτελέσματα που αποκλίνουν σημαντικά τόσο από την ανθρώπινη κρίση όσο και μεταξύ τους. Η «βεβαιότητα» που εκπέμπει ένα μοντέλο δεν αντικατοπτρίζει την ακρίβειά του (Bender et al., 2021; Bommasani et al., 2021).

Αυτό δεν σημαίνει ότι τα ΜΓΜ είναι άχρηστα, αλλά αποτυπώνουν ότι η χρήση τους απαιτεί κριτική στάση: επαλήθευση έναντι ανθρώπινης επισημείωσης, σύγκριση μεταξύ πολλαπλών συστημάτων, και επίγνωση των ορίων που θέτει η γλωσσική και πολιτισμική ιδιαιτερότητα κάθε εφαρμογής (Artstein & Poesio, 2008). Στο πεδίο της υπολογιστικής γλωσσολογίας και ιδίως για γλώσσες χαμηλών πόρων η αυθεντία δεν κατακτάται από την τεχνολογική πολυπλοκότητα ενός συστήματος, αλλά από την εξάσκηση, την διαφάνεια, την επαληθευσιμότητα και την κριτική αξιολόγηση των αποτελεσμάτων του (Riloff et al., 2013).

# Βιβλιογραφικές αναφορές

Amin, M., & Burghardt, M. (2023). A wide-coverage evaluative framework for figurative language detection. *Natural Language Engineering, 30*(1), 1–29. https://doi.org/10.1017/S1351324923000268

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics, 32*(6), 793–826. https://doi.org/10.1016/S0378-2166(99)00070-3

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bakagianni, J., Pouli, K., Gavriilidou, M., & Pavlopoulos, J. (2025). A systematic survey of natural language processing for the Greek language. Δημοσιεύθηκε 15 Ιουλίου 2025

Brown, Tom & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda & Agarwal, Sandhini & Herbert-Voss, Ariel & Krueger, Gretchen & Henighan, Tom & Child, Rewon & Ramesh, Aditya & Ziegler, Daniel & Wu, Jeffrey & Winter, Clemens & Amodei, Dario. (2020). Language Models are Few-Shot Learners. 10.48550/arXiv.2005.14165.

Chalamandaris, A., Tsiakoulis, P., Zervas, P., & Karabetsos, S. (2006). Language independent and data driven grapheme to phoneme conversion. *Proceedings of LREC 2006*, 121–126.

Γούτσος, Δ. (2020) *Γλώσσα. Κείμενο, ποικιλία, σύστημα*. Αθήνα: Κριτική. Δεύτερη έκδοση.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the*

*Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Culpeper, J., Haugh, M., & Kádár, D. Z. (Eds.). (2017). *The Palgrave handbook of linguistic (im)politeness*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-37508-7

Culpeper, J., Mackey, A., & Taguchi, N. (2017). *Second language pragmatics: From theory to research*. Routledge.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. *Proceedings of CoNLL 2010*, 107–116.

DeepSeek-AI. (2024). DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Frawley W. Norman Fairclough, Discourse and social change. Cambridge: Polity, 1992. Pp. vii + 259. *Language in Society*. 1993;22(3):421-424. doi:10.1017/S0047404500017309

Fuoil, M., Huang, W., Littlemore, J., Turner, S., & Wilding, E. (2026). Metaphor identification using large language models: A comparison of RAG, prompt engineering, and fine-tuning. *Applied Corpus Linguistics, 6*, Article 100204.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol, 15*(1–2), 5–27. https://doi.org/10.1080/10926488.2000.9678862

Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41–58). Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

Haber, J., & Poesio, M. (2020). It's not just about token-level accuracy: Annotation of diverse metaphors using crowdsourcing. *Proceedings of the 28th International Conference on Computational Linguistics*, 4256–4267.

Happé, F. G. E. (1993). Communicative competence and theory of mind in autism. *Cognition, 48*(2), 101–119. https://doi.org/10.1016/0010-0277(93)90039-X

Hovy, E., & Lavid, J. (2010). Towards a "science" of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation, 22*(1), 13–36.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of KDD '04*, 168–177. https://doi.org/10.1145/1014052.1014073

Huang, Y. (Ed.). (2017). *The Oxford handbook of pragmatics*. Oxford Academic. https://doi.org/10.1093/oxfordhb/9780199697960.001.0001

Jang, H., & Shin, H. (2020). It's not just sarcasm: Multilevel figurative language annotation. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6367–6374.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213.

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Κουτσός, Π., & Αναγνωστόπουλος, Χ. (2022). Ανάλυση συμπεριφοράς καταναλωτών στο ελληνικό ηλεκτρονικό εμπόριο. *Ελληνική Επιθεώρηση Μάρκετινγκ, 14*(2), 45–67.

Lakoff, G., & Johnson, M. (1980). *Metaphors we Live By*. University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G. (2010). Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, 117–125.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions* (2nd ed.). Cambridge University Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Markantonatou, S., Samaridi, N., Minos, M., & Sofianopoulos, S. (2019). Processing modern Greek with a view to MT. *Proceedings of the MT Summit XVII*, 113–123.

Maynard, D., & Greenwood, M. (2014). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 4238–4243.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Oprea, S. V., & Bâra, A. (2025). LLM-as-a-judge for sarcasm detection using supervised fine-tuning of transformers. *Journal of King Saud University – Computer and Information Sciences, 37*, 357. https://doi.org/10.1007/s44443-025-00379-7

Ortega Bueno, R. (2025). *New avenues in computational irony detection in social media* [Doctoral dissertation, Universitat Politècnica de València].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Περήφανος, Κ. (2019). *Ανάλυση Ελληνικών Σωμάτων Κειμένων με τη Χρήση Τεχνικών Μηχανικής Μάθησης: Υπολογιστική Αναπαράσταση της Ιδιολέκτου* [Διδακτορική διατριβή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών]. https://freader.ekt.gr/eadd/index.php?doc=45377

Pitenis, Z., Zampieri, M., & Ranasinghe, T. (2021). Offensive language identification in Greek. *Proceedings of RANLP 2021*, 1120–1128. https://doi.org/10.26615/978-954-452-072-4_126

Poesio, M., Chamberlain, J., & Kruschwitz, U. (2019). Crowdsourcing. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 247–276). Springer.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42–49.

Polikar, Robi. (2006). Polikar, R.: Ensemble based systems in decision making. IEEE Circuit Syst. Mag. 6, 21-45. Circuits and Systems Magazine, IEEE. 6. 21 - 45. 10.1109/MCAS.2006.1688199.

Potamias, R. A., Siolas, G., & Stafylopatis, A. (2019). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications, 32*, 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

Prokopidis, P., & Piperidis, S. (2020). A neural NLP toolkit for Greek. *Proceedings of the 11th Hellenic Conference on Artificial Intelligence*, 138–143. https://doi.org/10.1145/3411408.3411430

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 704–714.

Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics, 41*(4), 579–623. https://doi.org/10.1162/COLI_a_00233

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. In P. Cole (Ed.), *Radical pragmatics* (pp. 295–318). Academic Press.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.

Sperber, D., & Wilson, D. (2015). Beyond speaker's meaning. *Croatian Journal of Philosophy, 15*(44), 117–149.

Stowe, K., Utama, P., & Gurevych, I. (2022). IMPLI: Investigating NLI models over implied presuppositions and implicit language. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2634–2651. https://doi.org/10.18653/v1/2022.acl-long.182

Taguchi, Naoko & Culpeper, Jonathan & Mackey, Alison. (2018). Second Language Pragmatics: From Theory to Research. 10.4324/9781315692388.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 248–258.

Van Hee, C., Lefever, E., & Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. *Proceedings of the 12th International Workshop on Semantic Evaluation*, 39–50. https://doi.org/10.18653/v1/S18-1005

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention Is All You Need. arXiv: 1706.03762.

Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons. *Proceedings of ECAI 2010*, 765–770.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. *International Conference on Learning Representations*. https://openreview.net/forum?id=gEZrGCozdqR

Wilson, D., & Sperber, D. (2012). Explaining irony. In D. Wilson & D. Sperber, *Meaning and relevance* (pp. 123–145). Cambridge University Press.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

# Παράρτημα

Ο κώδικας python που χρησιμοποιήθηκε για την παρούσα εργασία:

```
#!/usr/bin/env python3
"""Classify text as ironic or not using Claude with structured Pydantic output."""

import argparse
import sys

import anthropic
import instructor
from enum import Enum
from pydantic import BaseModel, Field


class IronyLabel(str, Enum):
    IRONY = "irony"
    NOT_IRONY = "not_irony"


class IronyClassification(BaseModel):
    label: IronyLabel = Field(
        description="Classification label: 'irony' or 'not_irony'"
    )
    confidence: float = Field(
        ge=0.0, le=1.0, description="Confidence score between 0.0 and 1.0"
    )
    reasoning: str = Field(description="Brief explanation for the classification")
```

```
SYSTEM_PROMPT = (
    "You are an expert linguist specializing in detecting irony and sarcasm. "
    "Given a text, determine whether it is ironic (includes verbal irony, situational irony, "
    "and sarcasm) or not ironic. Always respond with a structured classification."
)

USER_TEMPLATE = (
    "Classify the following text as ironic or not ironic.\n\n"
    "Text: {text}\n\n"
    "Consider:\n"
    "- Verbal irony: saying the opposite of what is meant\n"
    "- Sarcasm: mocking or contemptuous irony\n"
    "- Situational irony: an outcome contrary to what was expected\n\n"
    "Provide your label, a confidence score, and brief reasoning."
)


def classify(
    text: str, client: instructor.Instructor, model: str
) -> IronyClassification:
    return client.messages.create(
        model=model,
        max_tokens=512,
        system=SYSTEM_PROMPT,
        messages=[{"role": "user", "content": USER_TEMPLATE.format(text=text)}],
        response_model=IronyClassification,
    )


def print_result(
    text: str, result: IronyClassification, show_text: bool = True
```

```
) -> None:
    if show_text:
        print(f"Text:       {text}")
    print(f"Label:      {result.label.value}")
    print(f"Confidence: {result.confidence:.2f}")
    print(f"Reasoning:  {result.reasoning}")


def main() -> None:
    parser = argparse.ArgumentParser(
        description="Classify text as ironic or not ironic using Claude."
    )
    parser.add_argument("text", nargs="?", help="Text to classify")
    parser.add_argument(
        "--file", "-f", help="File with texts to classify (one per line)"
    )
    parser.add_argument(
        "--model",
        "-m",
        default="claude-opus-4-5",
        help="Claude model to use (default: claude-opus-4-5)",
    )
    args = parser.parse_args()

    client = instructor.from_anthropic(anthropic.Anthropic())

    if args.file:
        with open(args.file, encoding="utf-8") as f:
            texts = [line.strip() for line in f if line.strip()]
        for text in texts:
            result = classify(text, client, args.model)
```

```
            print_result(text, result)
            print("-" * 60)

    elif args.text:
        result = classify(args.text, client, args.model)
        print_result(args.text, result, show_text=False)

    else:
        # Interactive mode
        print("Irony Classifier — enter text to classify (Ctrl+C to exit)")
        print("-" * 60)
        while True:
            try:
                text = input("Text: ").strip()
                if not text:
                    continue
                result = classify(text, client, args.model)
                print_result(text, result, show_text=False)
                print()
            except KeyboardInterrupt:
                print("\nDone.")
                sys.exit(0)


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

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.