ΘΕΜΑ: Ανίχνευση Κειμένων Γραμμένων από AI σε Ελληνικά & Αγγλικά Κείμενα

Στόχος της εργασίας είναι η δημιουργία ελληνικού συνόλου δεδομένων με κείμενα γραμμένα από ανθρώπους και από LLMs. Πάνω σε αυτό αξιολογείται συγκριτικά πόσο καλά σύγχρονα LLMs (π.χ. GPT, Gemini) ανιχνεύουν κείμενα που έχουν παραχθεί από AI (machine-generated text detection), σε σενάρια zero-shot και few-shot.

Α. ΕΠΙΣΚΟΠΗΣΗ ΤΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

ΒΗΜΑ 1 (ΜΕΧΡΙ 3/11/2026):

- Αναζήτηση βιβλιογραφίας στο scholar.google.com με λέξεις, προτάσεις-κλειδιά όπως:
  - "machine-generated text detection"
  - "AI-generated text detection multilingual"
  - "zero-shot detection of LLM-generated text"
  - "large language models as detectors of AI-generated text"
  - "Greek large language models"
- Παρουσίαση υπαρχουσών μελετών:
  - Παραδοσιακές προσεγγίσεις.
  - Fine-tuned Transformers (π.χ. RoBERTa, XLM-R) ως ταξινομητές.
  - Zero-shot στατιστικές μέθοδοι (DetectGPT, Fast-DetectGPT, Binoculars) και υδατογραφήματα (watermarking).
  - LLMs ως ανιχνευτές μέσω prompting (π.χ. GPT, Gemini).

Β. ΕΠΙΛΟΓΗ ΚΑΙ ΔΗΜΙΟΥΡΓΙΑ ΔΕΔΟΜΕΝΩΝ

ΒΗΜΑ 2 (ΜΕΧΡΙ 6/11/2026):

Σε αυτό το βήμα φτιάχνετε ένα αρχείο Excel ή CSV. Κάθε γραμμή του είναι ένα κείμενο, γραμμένο είτε από άνθρωπο είτε από AI σύστημα. Ακολουθήστε τα παρακάτω βήματα με τη σειρά.

Βήμα 2.1 – Μαζέψτε κείμενα γραμμένα από ανθρώπους

- Συλλέξτε ελληνικά κείμενα και, για σύγκριση, λιγότερα αγγλικά.
- Κάθε κείμενο να έχει μήκος περίπου 100–300 λέξεις (1–3 παράγραφοι).
- Τα κείμενα πρέπει να έχουν γραφτεί πριν από το τέλος του 2022. Τότε κυκλοφόρησε το ChatGPT, οπότε έτσι είμαστε σίγουροι ότι τα έγραψε άνθρωπος.
- Πάρτε κείμενα από τουλάχιστον 2 θεματικά πεδία (π.χ. εγκυκλοπαιδικά και νομικά).
- Πού θα τα βρείτε:
  - **Ελληνικά:** Ελληνική Wikipedia (Wikimedia Dumps), Greek Legal Code (Hugging Face), ειδησεογραφικά άρθρα και blogs πριν από το 2022.
  - **Αγγλικά:** Αγγλική Wikipedia (Wikimedia Dumps), ειδησεογραφικά άρθρα και blogs πριν από το 2022.
- Για τη Wikipedia, πάρτε την έκδοση του άρθρου πριν από το 2022, από το «Ιστορικό» του άρθρου ή από παλαιότερο dump.

Βήμα 2.2 – Φτιάξτε τα αντίστοιχα κείμενα AI

- Για κάθε ανθρώπινο κείμενο, ζητήστε από ένα LLM να γράψει ένα νέο κείμενο για το ίδιο θέμα, στην ίδια γλώσσα και με περίπου το ίδιο μήκος.
- Χρησιμοποιήστε το ίδιο prompt για όλα τα κείμενα, π.χ.: *«Γράψε ένα κείμενο περίπου 200 λέξεων με θέμα "[τίτλος του ανθρώπινου κειμένου]", σε ύφος άρθρου εγκυκλοπαίδειας.»*
- Χρησιμοποιήστε τουλάχιστον 3 διαφορετικά LLMs (π.χ. GPT, Gemini, Claude, Llama-Krikri). Μοιράστε τα κείμενα περίπου ισόποσα ανάμεσά τους.
- Αντιγράψτε την απάντηση του μοντέλου όπως είναι, χωρίς διορθώσεις.

Βήμα 2.3 – Βάλτε όλα τα κείμενα σε ένα αρχείο

Κάθε γραμμή του αρχείου είναι ένα κείμενο, με τις εξής στήλες:

id	κείμενο	ετικέτα	γλώσσα	πηγή ή μοντέλο	θεματικό πεδίο
1	«Η Κέρκυρα είναι νησί του Ιονίου…»	Ανθρώπινο	EL	Wikipedia	εγκυκλοπαιδικό
2	«Η Κέρκυρα, το πράσινο νησί…»	AI	EL	Gemini	εγκυκλοπαιδικό
3	«Corfu is a Greek island…»	Ανθρώπινο	EN	Wikipedia	εγκυκλοπαιδικό

- Συνολικά 500–1000 γραμμές, μισές ανθρώπινες και μισές AI.
- Παράδειγμα για 600 γραμμές: 200 ελληνικά και 100 αγγλικά ανθρώπινα κείμενα, συν τα 300 αντίστοιχα κείμενα AI.
- Το prompt που χρησιμοποιήσατε το γράφετε μία φορά στην εργασία σας.

Βήμα 2.4 – Κρατήστε λίγα κείμενα για τα παραδείγματα του few-shot

- Ξεχωρίστε 10 κείμενα ανά γλώσσα (5 ανθρώπινα και 5 AI). Θα τα χρησιμοποιήσετε μόνο ως παραδείγματα στο few-shot του Βήματος 3.
- Αυτά τα κείμενα δεν μπαίνουν στην αξιολόγηση.

Κανόνας: Μην χρησιμοποιείτε κείμενα συμφοιτητών ή άλλα προσωπικά κείμενα.

Προαιρετικά: μπορείτε να προσθέσετε και δύσκολες περιπτώσεις, όπως:

- κείμενα AI που έχουν παραφραστεί ή «εξανθρωπιστεί»
- κείμενα μεταφρασμένα από τα αγγλικά
- κείμενα με code-switching (μείξη EL/EN)

Γ. ΕΦΑΡΜΟΓΗ ΣΥΣΤΗΜΑΤΩΝ

ΒΗΜΑ 3 (ΜΕΧΡΙ 27/11/2026):

- Επιλογή LLMs ως ανιχνευτών: τουλάχιστον δύο (π.χ. GPT, Gemini).
- Ορισμός πειραματικών σεναρίων: Θα τρέξετε τα ίδια κείμενα σε δύο στησίματα: χωρίς παραδείγματα (zero-shot) και με λίγα ενδεικτικά παραδείγματα (few-shot).
- **Zero-shot:** Δώστε στο μοντέλο μόνο το κείμενο και σαφείς οδηγίες για την ετικέτα που πρέπει να επιστρέψει. Δεν βάζουμε κανένα παράδειγμα λύσης μέσα στο prompt. Ζητάμε σύντομη και καθαρή απάντηση (μόνο την ετικέτα και μία σύντομη αιτιολόγηση).
- **Few-shot (3–5 παραδείγματα):** Πριν από το κείμενο προς αξιολόγηση, δίνουμε 3–5 μικρά, αντιπροσωπευτικά παραδείγματα με τη σωστή ετικέτα. Τα παραδείγματα δεν πρέπει να προέρχονται από τα κείμενα ελέγχου. Φροντίζουμε να καλύπτουν και τις δύο ετικέτες.
- Προαιρετικά, για βελτίωση του σκορ: καλύτερος σχεδιασμός του prompt, προσεκτική επιλογή των few-shot παραδειγμάτων, συνδυασμός πολλών LLMs.
- Συλλογή αποτελεσμάτων για κάθε LLM και κάθε σενάριο.

Δ. ΑΞΙΟΛΟΓΗΣΗ

ΒΗΜΑ 4 (ΜΕΧΡΙ 8/12/2026):

- Αυτόματη αξιολόγηση με μετρικές:
  - Accuracy, Precision, Recall, F1-score (macro).
  - Ποσοστό ψευδώς θετικών στα ανθρώπινα κείμενα (πόσο συχνά ένας άνθρωπος «κατηγορείται» άδικα ότι έγραψε με AI).
  - Confusion matrix (για να δούμε πού μπερδεύεται ο ανιχνευτής).
- Με τη χρήση της βιβλιοθήκης scikit-learn.
- Ποιοτική ανάλυση:
  - Ανθρώπινα κείμενα που χαρακτηρίστηκαν AI (π.χ. τυπικό ή νομικό ύφος).
  - Πολύ σύντομα ή ασαφή κείμενα.
  - Παραφρασμένα ή μεταφρασμένα κείμενα.

Ε. ΣΥΓΚΡΙΣΗ ΣΥΣΤΗΜΑΤΩΝ

ΒΗΜΑ 5 (ΜΕΧΡΙ 15/1/2027):

- Πίνακας με συγκριτικά αποτελέσματα GPT, Gemini (ίσως και άλλων LLMs), σε zero-shot και few-shot, για ελληνικά και αγγλικά.
- Ανάλυση διαφορών:
  - Ποιο LLM είναι πιο «καχύποπτο» (χαρακτηρίζει πολλά ανθρώπινα κείμενα ως AI) και ποιο πιο «επιεικές».
  - Αναγνωρίζει ένα μοντέλο τα δικά του κείμενα; (π.χ. εντοπίζει το GPT πιο εύκολα κείμενα του GPT απ' ό,τι του Gemini;)
  - Κείμενα ποιων LLMs ανιχνεύονται πιο δύσκολα.
  - Ελληνικά έναντι αγγλικών.

ΣΤ. ΔΟΜΗ ΕΡΓΑΣΙΑΣ

ΒΗΜΑ 6 (ΜΕΧΡΙ ΤΗΝ ΗΜΕΡΟΜΗΝΙΑ ΤΩΝ ΕΞΕΤΑΣΕΩΝ ΤΟΥ ΜΑΘΗΜΑΤΟΣ):

Συντάξτε επιστημονική εργασία στα Αγγλικά (σαν τις εργασίες της βιβλιογραφίας) μέχρι 10 σελίδες το πολύ. Η εργασία θα περιλαμβάνει:

1. Θεωρητικό υπόβαθρο (ανίχνευση κειμένων AI + LLMs).
2. Περιγραφή dataset (τρόπος κατασκευής, στατιστικά).
3. Πειραματική διαδικασία (LLMs ως ανιχνευτές, zero-shot/few-shot).
4. Αποτελέσματα & γραφήματα (accuracy, F1).
5. Ανάλυση σφαλμάτων.
6. Συμπεράσματα: τι δουλεύει καλύτερα, τι όχι.
7. Μελλοντικές βελτιώσεις.
8. Βιβλιογραφία.

ΕΠΙΠΛΕΟΝ

- Η εργασία θα πραγματοποιηθεί σε ομάδες των τεσσάρων ατόμων.
- Η εργασία είναι υποχρεωτική και πιάνει το 50% στον τελικό βαθμό.