Πειραματισμός με Large Language Models (LLMs): Hallucinations & Ακολουθία Context

Στόχος της εργασίας είναι να κατανοήσετε στην πράξη:

- τι είναι τα hallucinations στα Large Language Models (LLMs),
- αν και πότε ένα LLM ακολουθεί ή αγνοεί το δοσμένο context,

Η εργασία δεν απαιτεί προγραμματισμό. Μπορεί να γίνει αποκλειστικά μέσω web εργαλείων (π.χ. ChatGPT, Gemini, Copilot κ.λπ.).

ΜΕΡΟΣ Α — Έλεγχος Hallucinations (Υποχρεωτικό)

Θα χρησιμοποιήσετε **τις παρακάτω 4 ερωτήσεις γνώσης**:

1. Ποια είναι η πρωτεύουσα της Αυστραλίας;
   → **Σωστή απάντηση:** Canberra
2. Τι σημαίνουν τα αρχικά CPU;
   → **Σωστή απάντηση:** Central Processing Unit
3. Στην ιατρική, τι σημαίνει ECG;
   → **Σωστή απάντηση:** Electrocardiogram
4. Ποια χρονιά έγινε για πρώτη φορά διαθέσιμο δημόσια το ChatGPT;
   → **Σωστή απάντηση:** 2022

Θα χρησιμοποιησετε τρια μοντελα πχ (Μπορειτε να χρησιμοποιησετε και αλλα):

- ChatGPT
- Gemini
- Copilot

Για **κάθε ερώτηση**, θα χρησιμοποιήσετε **δύο διαφορετικά prompts**:

- ❖ Prompt 1 (Απλό)

«Απάντησε στην παρακάτω ερώτηση με μία σύντομη πρόταση:
[ερώτηση]»

- ❖ Prompt 2 (Περιοριστικό)

«Εάν δεν είσαι 100% σίγουρος για την απάντηση, πρέπει να απαντήσεις μόνο: "Δεν γνωρίζω". Ερώτηση: [ερώτηση]»

- ➢ **Τι πρέπει να καταγράψετε**

Για κάθε ερώτηση και για κάθε prompt συμπληρώνετε τον παρακάτω πίνακα:

Μοντελο	Ερώτηση	Prompt	Απάντηση Μοντέλου	Σωστή; (Ναι/Όχι)	Hallucination; (Ναι/Όχι)

Παράδειγμα:

Ερώτηση	Model	Prompt	Answer	Correct	Hallucination
Q1	ChatGPT	Simple	…	Yes	No
Q1	Gemini	Simple	…	Yes	No
Q1	Copilot	Simple	…	Yes	No
Q1	ChatGPT	Constrained	…	Yes	No
…	…	…	…	…	…

Hallucination = απάντηση που δίνεται με σιγουριά αλλά είναι λανθασμένη.

Στο τέλος γράφετε **5–6 γραμμές σχόλιο**, π.χ.:

- Παρατηρήσατε λιγότερα λάθη με το δεύτερο prompt;
- Είπε ποτέ το μοντέλο «Δεν γνωρίζω»;
- Σας έκανε κάτι εντύπωση;

ΜΕΡΟΣ Β — Έλεγχος Ακολουθίας Context (Υποχρεωτικό)

Το Μέρος Β να πραγματοποιηθεί **σε ένα μόνο από τα παραπάνω μοντέλα της επιλογής σας**.

Χρησιμοποιήστε το παρακάτω φανταστικό context:

Context:

«Η πόλη Avenport είναι η πρωτεύουσα ενός φανταστικού νησιωτικού κράτους στον Ειρηνικό Ωκεανό.
Η πρωτεύουσα της Avenport ονομάζεται Lunaris.»

Ερώτηση:

«Με βάση το παραπάνω κείμενο, ποια είναι η πρωτεύουσα της Avenport;»

Θα χρησιμοποιήσετε **δύο prompts**:

- **Prompt 3 (Χαλαρό)**

«Διάβασε το παρακάτω κείμενο και απάντησε στην ερώτηση.»

- **Prompt 4 (Αυστηρό)**

«ΠΡΕΠΕΙ να απαντήσεις ΜΟΝΟ βάσει του παρακάτω κειμένου.
Αγνόησε οποιαδήποτε πραγματική γνώση.
Αν δεν υπάρχει απάντηση στο κείμενο, να πεις: "Δεν γνωρίζω".»

➢ **Τι πρέπει να καταγράψετε**

Συμπληρώνετε τον παρακάτω πίνακα:

Μοντέλο	Prompt	Απάντηση Μοντέλου	Ακολούθησε το context; (Ναι/Όχι)	Σχόλιο

Και γράφετε **5–6 γραμμές συμπέρασμα**, π.χ.:

- Ακολούθησε το context ή το αγνόησε;
- Βοήθησε το αυστηρό prompt;

Τι πρέπει να παραδώσετε

Ένα **PDF (1–2 σελίδες)** που να περιλαμβάνει:

1. Τους πίνακες του Μέρους Α
2. Τον πίνακα του Μέρους Β
3. Τα σύντομα σχόλια – συμπεράσματα
4. **Screenshots επιτρέπονται αλλά δεν είναι υποχρεωτικά.**