ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Πτυχιακή Εργασία *2025-ΠΕ492*

Κατασκευή ενός RAG Chatbot για Ξενοδοχειακή Υποστήριξη

Νικόλαος Λυμπιτάκης

Επιβλέπων: Μηνάς Δασυγένης
Μέλη ΕΚ: Χρήστος Κούτσικας, Βασίλειος Στεφανής

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Δομή Παρουσίασης

- **Εισαγωγή & Κίνητρο** — Πρόβλημα & αναγκαιότητα λύσης
- **Θεωρητικό Υπόβαθρο** — Κατηγορίες chatbot, αρχιτεκτονική RAG
- **Τεχνολογίες υλοποίησης**— Python, LangChain, FAISS, DeepSeek, Streamlit
- **Σχεδιασμός Συστήματος** — Προδιαγραφές & Use Cases
- **Υλοποίηση** — Αρχιτεκτονική, Βάση Γνώσης, ReAct Agent,Caching (good_examples), Διεπαφή Χρήστη-UI
- **Αξιολόγηση** — Μεθοδολογία LLM-as-a-Judge & αποτελέσματα
- **Συμπεράσματα & Μέλλον** — SWOT, μελλοντικές επεκτάσεις
- **Επίδειξη** — Live demo του chatbot

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Εισαγωγή & Κίνητρο

- **Προβλήματα Παραδοσιακής Εξυπηρέτησης:**
  - Υψηλό λειτουργικό κόστος (μισθοί, εκπαίδευση)
  - Περιορισμένες ώρες διαθεσιμότητας
  - Αδυναμία κλιμάκωσης σε ώρες αιχμής
  - Ανομοιόμορφη ποιότητα εξυπηρέτησης
- **RAG Chatbot ως Λύση**:
  - 24/7 διαθεσιμότητα, άμεση απόκριση
  - Μείωση κόστους ~20%
  - Κλιμακώσιμο σε πολλές ταυτόχρονες συνομιλίες
  - Συνεπής ποιότητα πληροφορίας
  - Πολυγλωσσική υποστήριξη

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Θεωρητικό Υπόβαθρο — Κατηγορίες Chatbot

- Rule-Based — Κανόνες(if-else) & λέξεις κλειδιά. Απλά αλλά περιορισμένα.
- NLP-Based — Επεξεργασία φυσικής γλώσσας, πρόθεση & οντότητες. Καλύτερη κατανόηση.
- Generative (LLMs) — Παράγουν κείμενο. Φυσική επικοινωνία αλλά με hallucinations.
- RAG Chatbot — LLM + εξωτερική βάση γνώσης. Ακρίβεια & επικαιρότητα.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Αρχιτεκτονική RAG

- **Στάδιο 1: Σύστημα Ανάκτησης (The Retriever)**
  - Βήμα 1: Έγγραφα → Chunking → Embedding model → vector αναπαραστάσεις (διανύσματα) → αποθήκευση σε FAISS vector store
  - Βήμα 2: Ερώτηση χρήστη → Embedding model → vector αναπαραστάση → αναζήτηση ομοιότητας στο FAISS → retrieved context
- **Στάδιο 2: Σύστημα Παραγωγής (The Generator)**
  - Βήμα 1: Augmentation: Ερώτηση + retrieved context → augmented prompt
  - Βήμα 2: Τεκμηριωμένη Παραγωγή: LLM χρησιμοποιεί retreived context ως εξωτερική γνώση → Παράγει τελική συνεκτική & πραγματικά ορθή απάντηση

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Τεχνολογίες Υλοποίησης

- Python — Κύρια γλώσσα προγραμματισμού
- LangChain — Framework διαχείρισης ροής & ReAct Agent
- FAISS (Facebook AI Similarity Search) — βιβλιοθήκη για αναζήτηση ομοιότητας/αποθήκευση embeddings
- DeepSeek — Γλωσσικό μοντέλο (LLM)
- Streamlit — Γραφικό περιβάλλον UI
- SQLite — Βάση δεδομένων δωματίων & κρατήσεων

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

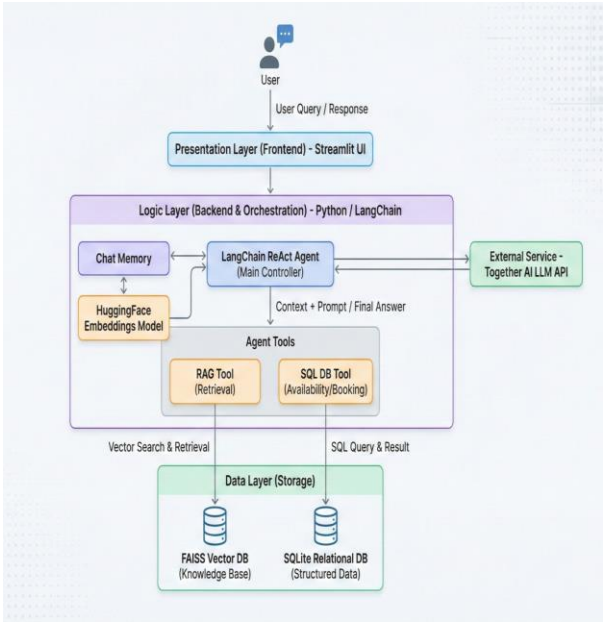
Σχεδιασμός — Προδιαγραφές & Use Cases

- **Λειτουργικές προδιαγραφές:**
  - Απάντηση σε ερωτήσεις FAQs & πολιτικές ξενοδοχείου
  - Αναζήτηση κοντινών σημείων ενδιαφέροντος
  - Αναζήτηση κριτικών
  - Έλεγχος διαθεσιμότητας δωματίων (SQLite)
  - Πολυγλωσσική υποστήριξη (ΕΛ / EN)
  - Μηχανισμός feedback (👍 / 👎) & Μνήμη συνομιλίας
- **Μη Λειτουργικές:**
  - Ακρίβεια απαντήσεων ≥ 80%
  - Χρόνος απόκρισης ≤ 15 δευτ.
- **Use Cases:**
  - UC1: Πολυγλωσσική επικοινωνία
  - UC2: Ερώτηση για παροχές, Έλεγχος διαθεσιμότητας
  - UC3: Πληροφορίες κοντινών σημείων ενδιαφέροντος & κριτικών
  - UC4: Feedback χρήστη

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Υλοποίηση — Αρχιτεκτονική Συστήματος

- **Επίπεδο 1 —Παρουσίασης (Presentation Layer):**
  - Streamlit
  - Επιλογή γλώσσας (sidebar), Feedback buttons
- **Επίπεδο 2 —Λογικής Chatbot (Logic Layer) :**
  - ReAct Agent (LangChain + DeepSeek)
  - Tool Selection (FAQ tool, Reviews Tool, POI tool, SQL tool)
  - Παραγωγή απάντησης
- **Επίπεδο 3 —Δεδομένων (Data Layer):**
  - FAISS vector stores (FAQ, Reviews, POI, Good_Examples)
  - SQLite database (δωμάτια, κρατήσεις, conversation logs)
  - Embeddings: all-MiniLM-L6-v2 model



Διάγραμμα αρχιτεκτονικής

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Βάση Γνώσης του Chatbot

- **5 πηγές δεδομένων που τροφοδοτούν τον agent:**
  - FAQs: Πολιτικές, παροχές, ώρες σε CSV → FAISS
  - Κριτικές Πελατών: Αξιολογήσεις σε CSV → FAISS
  - Σημεία Ενδιαφέροντος (POI): Αξιοθέατα, δραστηριότητες κτλ σε CSV → FAISS
  - SQLite Database: hotel_rooms, bookings, conversation_turns(logs)
  - Good Examples: Cached απαντήσεις με feedback=καλό → FAISS
- **Ροή δεδομένων:**
  - CSV Datasets (FAQs, Reviews, POI)
  - → Embeddings: all-MiniLM-L6-v2 model
  - → FAISS: Αποθήκευση σε vector stores

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

System Prompt template

- **Δομή Prompt Template:**
  - Ρόλος: Εικονικός βοηθός του 'The Grand Resort Hotel'
  - Κανόνες: Απαντήσεις μόνο από Tools & Knowledge Base
  - Αποτροπή παραισθήσεων: Άρνηση απάντησης όταν δεν υπάρχουν διαθέσιμα δεδομένα
  - Γλώσσα: Αυτόματη προσαρμογή στη γλώσσα χρήστη (ΕΛ/EN)
  - Conversation Memory: Ενσωμάτωση chat_history στο prompt
  - Ύφος: Επαγγελματικό & πειστικό
  - Χρήση good examples vector store: cached ερωτήσεις με feedback=καλό

ReAct Agent

- **ReAct (Reasoning and Acting) Agent του framework LangChain:**

- **Διαθέσιμα Tools του Agent:**
  - get_faq_answer() → Αναζήτηση στα FAQs (FAISS)
  - get_review_summary() → Περίληψη κριτικών πελατών (FAISS)
  - find_points_of_interest() → Σημεία Ενδιαφέροντος (FAISS)
  - get_available_rooms() → Διαθεσιμότητα δωματίων (SQL)
  - get_room_info() → Πληροφορίες & τιμές δωματίων (SQL)

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Μηχανισμός Good Examples (Caching)

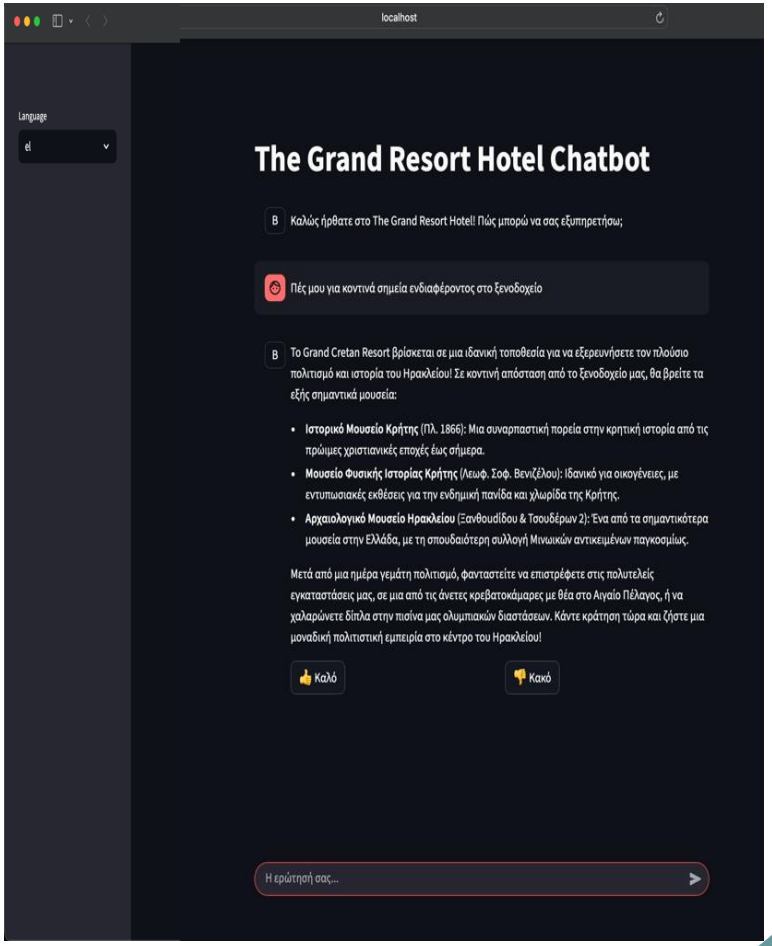
- Νέα / σύνθετη ερώτηση (χωρίς caching):
  - Thought → Action(επιλογή tool) → Observation → Επανάληψη → Final Answer (~15-25 δευτ.)
- Cached ερώτηση (good_examples vector store):

  Similarity search στο good_examples vector store →Παρόμοιο παράδειγμα βρέθηκε → Bypass Tool calls → Άμεση τελική απάντηση < 15 δευτ.
- Αποτέλεσμα: Σημαντική μείωση latency σε επαναλαμβανόμενες ερωτήσεις με καλό feedback.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Διεπαφή Χρήστη — Streamlit UI

- Welcome Message — Αυτόματο μήνυμα εκκίνησης
- Επιλογέας γλώσσας— Sidebar επιλογή ΕΛ/ΕΝ
- Μνήμη προηγούμενων συνομιλιών— Κλάση *ConversationBufferWindowMemory* του LangChain με k=5
- Feedback — 👍 / 👎 ανά απάντηση chatbot



ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Αξιολόγηση — Μεθοδολογία LLM-as-a-Judge

- **Διαδικασία αξιολόγησης :**
  - 30 Test Scenarios (15 ΕΛ + 15 ΕΝ ερωτήσεις)
  - Το DeepSeek δεν βαθμολογεί μόνο την απάντηση του chatbot, αλλά τη συγκρίνει με την αναμενόμενη σωστή απάντηση (Ground Truth), ώστε η αξιολόγηση να είναι αντικειμενική.
- **Μετρικές Αξιολόγησης:**
  - Accuracy — Ακρίβεια πληροφοριών (0-5)
  - Relevance — Συνάφεια με την ερώτηση (0-5)
  - Persuasiveness — Πειστικότητα (marketing ύφος) (0-5)
  - Faithfulness — Πιστότητα - Αποφυγή hallucinations (0=χωρίς παραισθήσεις)
  - Latency — Χρόνος απόκρισης (δευτ.)

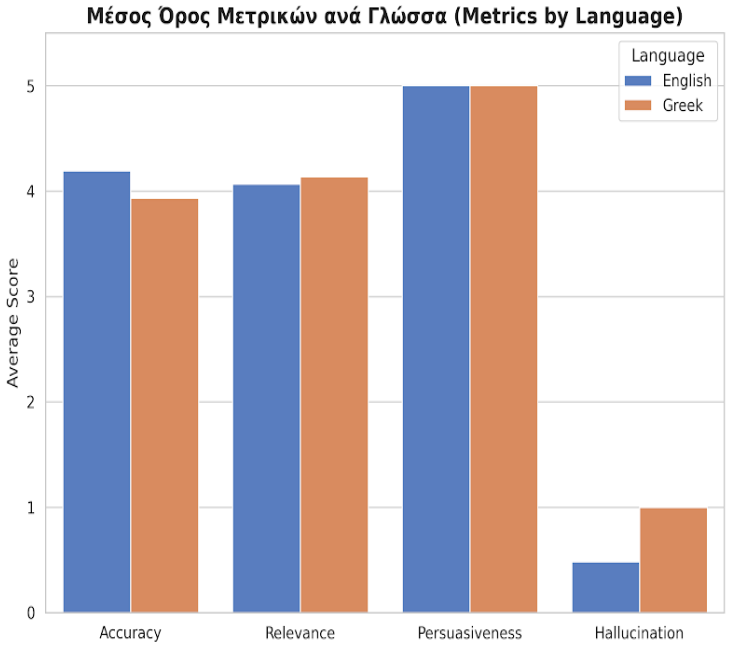
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Αποτελέσματα Αξιολόγησης

- Accuracy: 4.07/5 (81.4%) — Πάνω από στόχο 80%
- Relevance: 4.10/5 (82%)
- Persuasiveness: 5/5 (100%)
- Hallucination score: 0.73/5 (14.6%)
- Latency: 14.51 δευτ. μέσος χρόνος — Εντός ορίου 15 δευτ.
- Σύγκριση Αγγλικά / Ελληνικά:

  Αγγλικά:
  - Hallucination score = 0.5
  - Latency ~10 δευτ,

  Ελληνικά:
  - Hallucination score = 1.0>0.5 λόγω ότι η βάση γνώσης είναι στα αγγλικά
  - Latency ~19 δευτ. λόγω μετάφρασης



Διάγραμμα μέσου όρου μετρικών ανά γλώσσα

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Συμπεράσματα — Ανάλυση SWOT

- **Δυνατά Σημεία:**
  - RAG αρχιτεκτονική — ελάχιστες παραισθήσεις
  - Υβριδική διαχείριση δεδομένων (FAISS + SQLite), Good Examples caching
  - Πολυγλωσσική υποστήριξη
- **Αδυναμίες:**
  - Αυξημένο latency στα Ελληνικά
  - Εξάρτηση από DeepSeek API
  - Χειροκίνητη ενημέρωση vector stores
- **Ευκαιρίες:**
  - Σύνδεση με συστήματα κρατήσεων (Booking Engines)
  - Ενσωμάτωση σε WhatsApp / Messenger
- **Απειλές:**
  - Ασφάλεια και προστασία δεδομένων (GDPR)
  - Ανταγωνισμός
  - Μεταβολές κόστους API

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Μελλοντικές Βελτιώσεις & Επεκτάσεις

- **Streaming Αποκρίσεων:**
  - Σταδιακή εμφάνιση απάντησης — μείωση αντιληπτού χρόνου αναμονής
- **Τοπικά Μοντέλα (Local LLMs):**
  - Llama 3 ή Mistral σε ιδιόκτητους servers
  - Προστασία δεδομένων & μείωση κόστους API
- **Αυτοματοποιημένη Ενημέρωση Βάσης γνώσης:**
  - Αυτόματη ενημέρωση και ενσωμάτωση Vector Stores (π.χ. από νέα reviews)
  - Χωρίς χειροκίνητη εκτέλεση scripts

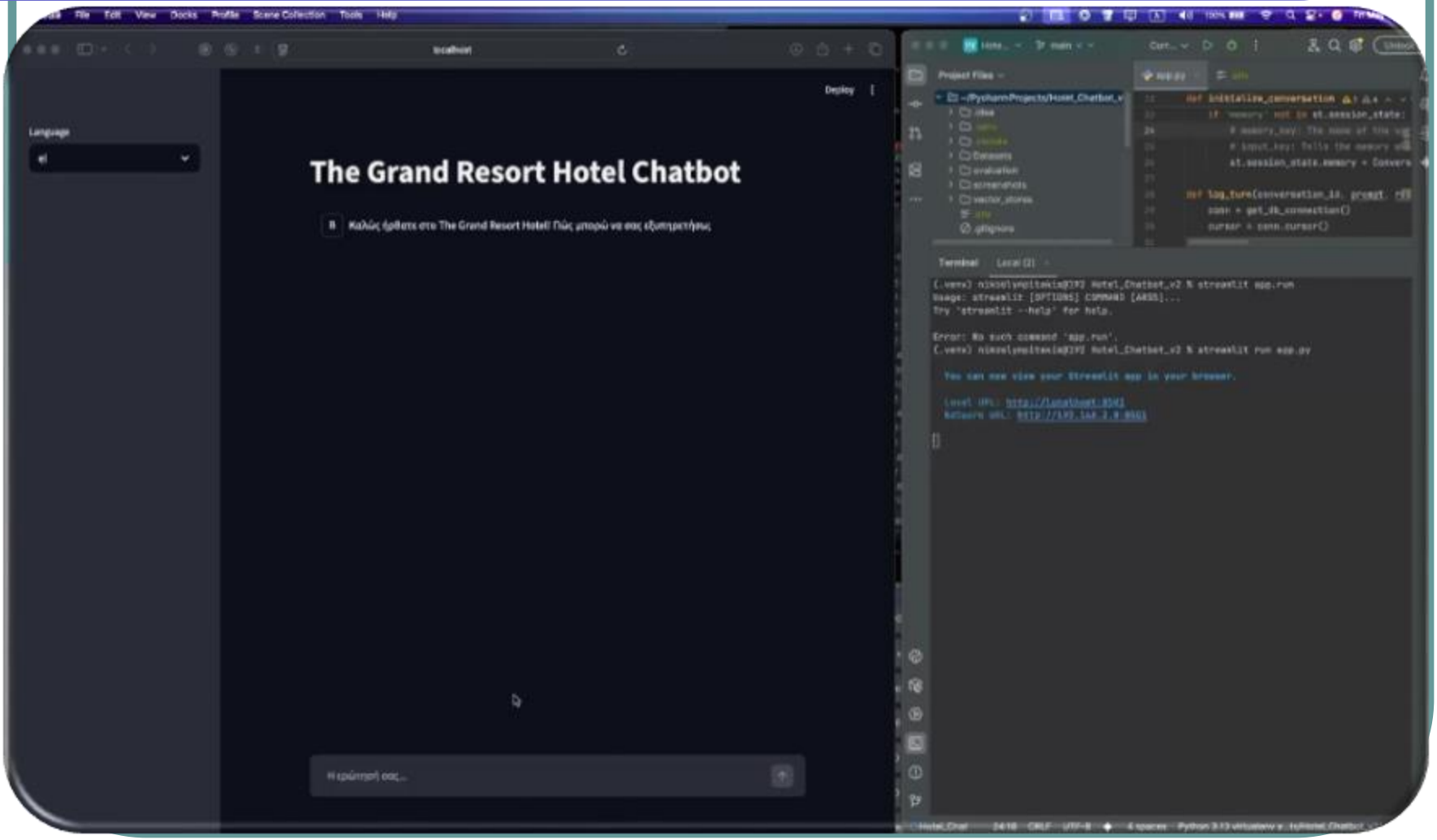
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Συμπεράσματα

- **RAG ως κατάλληλη αρχιτεκτονική:**
  - LLM + εξωτερική βάση γνώσης = ιδανικό για επιχειρησιακές εφαρμογές με αξιόπιστες και τεκμηριωμένες απαντήσεις
- **Υψηλή ποιότητα απαντήσεων:**
  - Accuracy 81.4%, Persuasiveness 100%, Hallucination score ≈ 0.73/5= 14.6%
- **Αποδοτική απόκριση:**
  - Μέσος χρόνος 14.51 δευτ. (εντός στόχου)
- **Πρακτική αξία για τον ξενοδοχειακό κλάδο:**
  - 24/7 εξυπηρέτηση, μείωση κόστους, πολυγλωσσική υποστήριξη

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Επίδειξη Λειτουργίας



ΕΛΛΗΝΙΚΟ
ΑΝΟΙΚΤΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ

Ευχαριστώ για την προσοχή σας!

Ερωτήσεις;