Education with Language Models: Analyzing Uncertainty Estimation Techniques

Dafni Tziakouri

July 11, 2024

Supervisor:
Dr. Jordi Vitrià

Outline

Introduction

Educational Tutor

Uncertainty Estimation

Experiments

Conclusions

Introduction

Motivation

Are LLMs ready to be used in education?

How uncertain are the output answers from LLMs?

Background

- What is a Large Language Model?
  - Categories of LLMs:
    1. Generic or raw
    2. Instruction-tuned
    3. Dialog-tuned
- How do LLMs work?
  1. Training
  2. Fine-tuning
  3. Prompt-tuning

Background

- What is a Large Language Model?
    - Categories of LLMs:
        1. Generic or raw
        2. Instruction-tuned
        3. Dialog-tuned
- How do LLMs work?
    1. Training
    2. Fine-tuning
    3. Prompt-tuning

GPT models

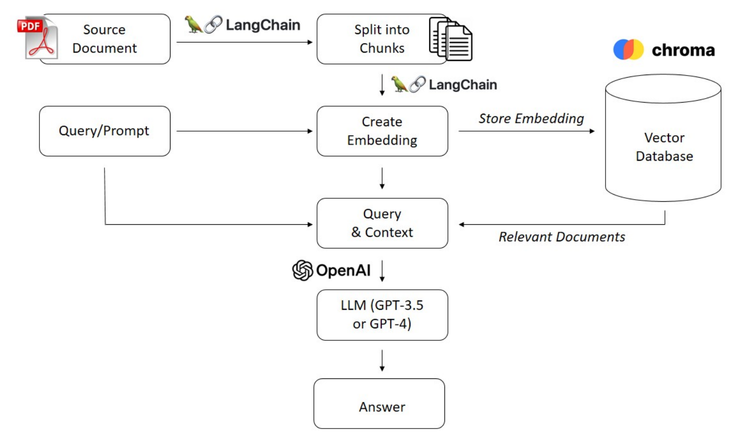
- GPT-3 and GPT-4
- Reinforcement Learning with Human Feedback (RLHF) / Rule-Based Reward Models (RBRMs)
- Transformer architecture
- Parameters: 175 billion (GPT-3) and 1.76 trillion (GPT-4)
- Limitations:
    1. Incapable of autonomous learning
    2. Hallucinations
    3. Inability to replicate human traits
    4. Challenges in source citation
    5. Bias, discrimination and stereotype

GPT models

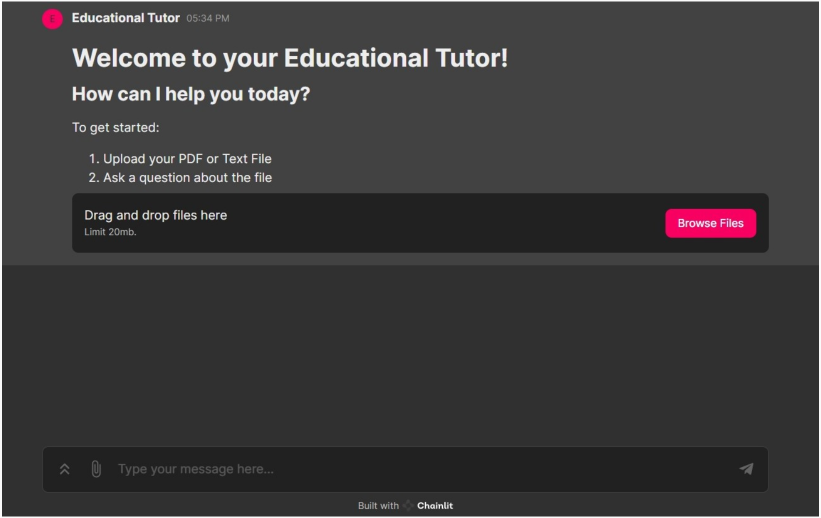
- GPT-3 and GPT-4
- Reinforcement Learning with Human Feedback (RLHF) / Rule-Based Reward Models (RBRMs)
- Transformer architecture
- Parameters: 175 billion (GPT-3) and 1.76 trillion (GPT-4)
- Limitations:
    1. Incapable of autonomous learning
    2. Hallucinations
    3. Inability to replicate human traits
    4. Challenges in source citation
    5. Bias, discrimination and stereotype

Educational Tutor

Pipeline


Source Document
LangChain
Split into Chunks
chroma
LangChain
Query/Prompt
Create Embedding
Store Embedding
Vector Database
Query & Context
Relevant Documents
OpenAI
LLM (GPT-3.5 or GPT-4)
Answer

Examples



Educational Tutor 05:34 PM

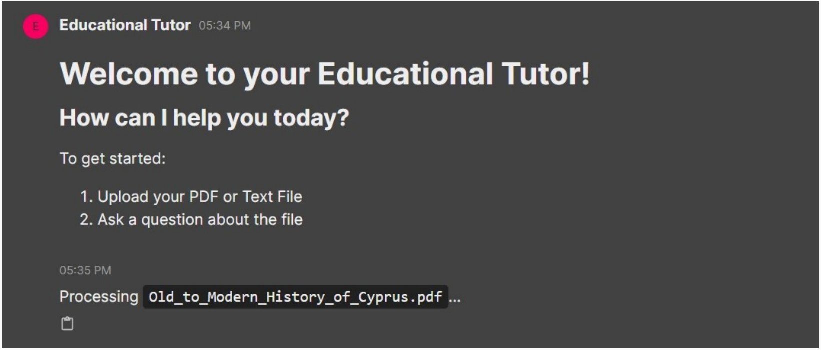
Welcome to your Educational Tutor!

How can I help you today?

To get started:

1. Upload your PDF or Text File
2. Ask a question about the file

Examples



Educational Tutor 05:34 PM

Welcome to your Educational Tutor!

How can I help you today?

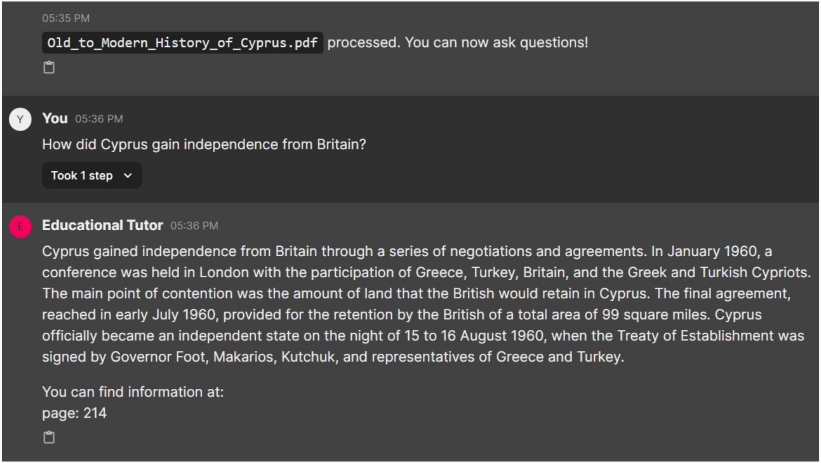
To get started:

1. Upload your PDF or Text File
2. Ask a question about the file

05:35 PM

Processing `Old_to_Modern_History_of_Cyprus.pdf`...

Examples



05:35 PM

`Old_to_Modern_History_of_Cyprus.pdf` processed. You can now ask questions!

You 05:36 PM

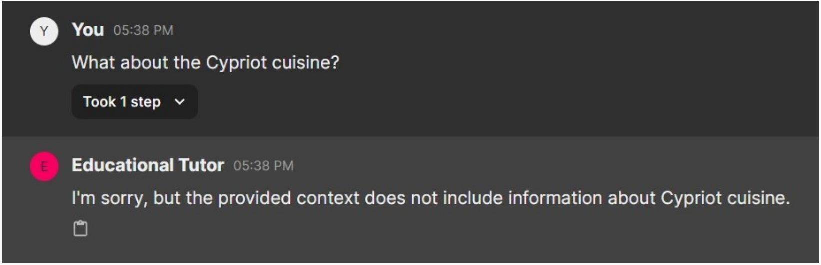
How did Cyprus gain independence from Britain?

Educational Tutor 05:36 PM

Cyprus gained independence from Britain through a series of negotiations and agreements. In January 1960, a conference was held in London with the participation of Greece, Turkey, Britain, and the Greek and Turkish Cypriots. The main point of contention was the amount of land that the British would retain in Cyprus. The final agreement, reached in early July 1960, provided for the retention by the British of a total area of 99 square miles. Cyprus officially became an independent state on the night of 15 to 16 August 1960, when the Treaty of Establishment was signed by Governor Foot, Makarios, Kutchuk, and representatives of Greece and Turkey.

You can find information at:
page: 214

Examples



What about the Cypriot cuisine?

I'm sorry, but the provided context does not include information about Cypriot cuisine.

Uncertainty Estimation

Uncertainty in technical terms

- Definition: Quantifying the degree of confidence in the predictions made by ML model.
- Methodology and Estimators:
  - **LM-Polygraph**
  - Focus on Graph Laplacian Eigenvalue Sum, Degree Matrix and Lexical Similarity.
- Framework:
  1. For a given input x, generate K response samples $y_1, \ldots, y_K$.
  2. Compute a $K \times K$ similarity matrix S between responses, where $S_{ij} = s(y_i, y_j)$ for some similarity score s (Jaccard score or Natural Language Inference score).
  3. Based on the similarity matrix S, compute the final uncertainty score.

Uncertainty in technical terms

- Definition: Quantifying the degree of confidence in the predictions made by ML model.
- Methodology and Estimators:
  - **LM-Polygraph**
  - Focus on Graph Laplacian Eigenvalue Sum, Degree Matrix and Lexical Similarity.
- Framework:
  1. For a given input x, generate K response samples $y_1, \dots, y_K$.
  2. Compute a $K \times K$ similarity matrix S between responses, where $S_{ij} = s(y_i, y_j)$ for some similarity score s (Jaccard score or Natural Language Inference score).
  3. Based on the similarity matrix S, compute the final uncertainty score.

Measuring answer similarities

Natural Language Inference (NLI):

1. Pair of responses: y and y'
2. Classification model: DeBERTa-large
3. Probabilities: $\hat{p}_{\text{entail}}(y, y')$ and $\hat{p}_{\text{contra}}(y, y')$
4. Measure of similarity between sequences y and y':

$$s_{\text{entail}}(y, y') = \hat{p}_{\text{entail}}(y, y')$$

or

$$s_{\text{contra}}(y, y') = 1 - \hat{p}_{\text{contra}}(y, y')$$

Estimating uncertainty from similarities

Sum of eigenvalues of the graph Laplacian:

1. Pairwise similarities: $s_{\text{entail}}(y_{j_1}, y_{j_2})$ and $s_{\text{entail}}(y_{j_2}, y_{j_1})$ for $j_1, j_2 = 1, \ldots, K$
2. Weighted matrix: $S_{j_1,j_2} = (s_{\text{entail}}(y_{j_1}, y_{j_2}) + s_{\text{entail}}(y_{j_2}, y_{j_1}))/2$
3. Symmetric normalized graph Laplacian

$$L := I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}} \tag{1}$$

$$D_{j_1,j_2} = \begin{cases} \sum_{j_2' \in [K]} S_{j_1,j_2'} & \text{if } j_1 = j_2 \\ 0 & \text{if } j_1 \neq j_2 \end{cases} \tag{2}$$

4. The uncertainty can be defined with $\lambda_1 < \ldots < \lambda_K$, the eigenvalues of L:

$$U_{\text{EigV}} = \sum_{k=1}^{K} \max(0, 1 - \lambda_k) \tag{3}$$

Estimating uncertainty from similarities

The degree matrix: elements of the diagonal D are sums of similarities between the given answer and the other answers,

$$U_{\text{Deg}}(x) = 1 - \text{trace}(D)/K$$

Lexical similarity: computes how similar two words or phrases are based on their meaning. It iterates over all responses and calculates the average similarity score with other answers. Metrics: ROUGE-1, ROUGE-2, ROUGE-L, etc.

Experiments

Experiments

LM-Polygraph

- **1st Case:** Lexical Similarity with the GPT-3.5 model.
- **2nd Case:** Lexical Similarity with the GPT-4 model.
- **3rd Case:** Graph Laplacian Eigenvalue Sum with the GPT-3.5 model.
- **4th Case:** Graph Laplacian Eigenvalue Sum with the GPT-4 model.

GLES: 19 questions
LS: 129 questions

Experiments

LM-Polygraph

- **1st Case:** Lexical Similarity with the GPT-3.5 model.
- **2nd Case:** Lexical Similarity with the GPT-4 model.
- **3rd Case:** Graph Laplacian Eigenvalue Sum with the GPT-3.5 model.
- **4th Case:** Graph Laplacian Eigenvalue Sum with the GPT-4 model.

GLES: 19 questions
LS: 129 questions

Research Questions

- **Research Question 1:** Which model provides the best uncertainty values for each method?
- **Research Question 2:** Does the Graph Laplacian Eigenvalue Sum method rank the uncertainty estimations in the same way as the Lexical Similarity method?
- **Research Question 3:** How reliable are the uncertainty estimations for each of the cases?

Results: Research Question 1

Table 1: Sample of uncertainty scores for GPT-3.5 and GPT-4 models with Lexical Similarity.

Input Prompt	Output Answer	Uncertainty with GPT-3.5	Uncertainty with GPT-4
tell me about the international and national arenas in Cyprus	Cyprus is a Mediterranean island country with a complex and unique international and national arena...	0.3512994	0.3369823
How did the european union react when cyprus was divided along ethnic lines in 1964 due to EOKA terror campaign?	The European Union did not exist in 1964, so it did not have a direct response to the division of Cyprus along ethnic lines due to the EOKA terror campaign...	0.3844868	0.3452402
Have the greek Cypriots always constituted a majority in Cyprus?	Yes, the Greek Cypriots have always constituted a majority in Cyprus...	0.4027987	0.3168868
What about the bronze time in cyprus	The Bronze Age in Cyprus lasted from around 2500 BC to 1050 BC...	0.3751035	0.3233221

Results: Research Question 1

Table 1: Sample of uncertainty scores for GPT-3.5 and GPT-4 models with Graph Laplacian Eigenvalue Sum.

Input Prompt	**Output Answer**	**Uncertainty with GPT-3.5**	**Uncertainty with GPT-4**
tell me about the international and national arenas in Cyprus	In the international arena, Cyprus is a member of numerous organizations and agreements, ...	3.7948324	6.0480381
How did the european union react when cyprus was divided along ethnic lines in 1964 due to EOKA terror campaign?	The European Union did not exist in 1964. The European Economic Community (EEC), which was the predecessor...	4.2764597	3.2875223
Have the greek Cypriots always constituted a majority in Cyprus?	Yes, the Greek Cypriots have always constituted a majority in Cyprus...	3.8615933	5.1233688
What about the bronze time in cyprus	The Bronze Age in Cyprus lasted from around 2300 to 1050 BC...	2.2675932	3.0284007

Results: Research Question 2

Table 2: Sample of ranked uncertainty scores for GPT-3.5 model with Lexical Similarity.

Input Prompt	Uncertainty with GPT-3.5
What were the major events and developments during the Byzantine rule of Cyprus?	0.2932287
What were some notable archaeological finds from Roman Cyprus?	0.2986870
What were the impacts of Ottoman rule on Cyprus, and how did the island adapt to its new rulers?	0.2987658
What were the outcomes of the Zurich-London agreements?	0.3020848
Can you give me suggested question that I might use in an exam for that period?	0.3143475

Results: Research Question 2

Table 2: Sample of ranked uncertainty scores for GPT-3.5 model with Graph Laplacian Eigenvalue Sum.

Input Prompt	Uncertainty with GPT-3.5
What about the bronze time in cyprus	2.2675932
What was the Greek Junta?	2.3972515
What evidence exists of Mycenaean influence on Cyprus?	2.5636839
How did Cyprus fare during the Bronze Age?	2.7019519
How did Cyprus come under Egyptian rule during the New Kingdom period?	2.9013555

Results: Research Question 2

Table 2: Sample of ranked uncertainty scores for GPT-4 model with Lexical Similarity.

Input Prompt	Uncertainty with GPT-4
Can you give me suggested question that I might use in an exam for that period?	0.1942356
What were some notable archaeological finds from Roman Cyprus?	0.1997253
How did Cyprus come under Egyptian rule during the New Kingdom period?	0.2116466
What were the challenges faced by Cyprus in implementing EU laws and regulations after accession?	0.2154664
How has the economy of Cyprus evolved over time?	0.2373326

Results: Research Question 2

Table 2: Sample of ranked uncertainty scores for GPT-4 model with Graph Laplacian Eigenvalue Sum.

Input Prompt	Uncertainty with GPT-4
What evidence exists of Mycenaean influence on Cyprus?	1.6800014
What role did Cyprus play in the trade networks of the ancient Mediterranean?	2.0429651
Give me questions about the Classical period in cyprus	2.1191324
What were the causes of the Turkish invasion in Cyprus?	2.9194977
What about the bronze time in cyprus	3.0284007

Results: Research Question 3

Table 3: Mean Uncertainty Scores for Each Case: Comparison of First and Second Datasets

Estimator	**Model**	**Mean for first dataset**	**Mean for second dataset**
Lexical similarity (LS)	GPT-3.5	0.4123847	0.4088312
Lexical similarity (LS)	GPT-4	0.3275486	0.3309661
Graph Laplacian (EVL)	GPT-3.5	3.7339862	4.2106864
Graph Laplacian (EVL)	GPT-4	3.9467708	3.5442523

Conclusions

Experiments conclusions

- Unstable and unbalanced results.
- Can not decide which model is better.
- Calculating the uncertainty of the **output answers** of LLMs may **not be efficient**.
- **Input prompt**, not the ideal for obtaining **reliable results** with black-box LLMs.
- Uncertainty, an **additional limitation** of LLMs.

General conclusions

- LLMs **impressive performance** in text generation tasks.
- GPT models should **not be** completely **trusted** for educational purposes.
- Very **limited** knowledge about estimating uncertainty in black-box models.
- **Research** in this field must **continue**.

Thank you for your attention! Any Questions?

UNIVERSITAT DE
BARCELONA