Harnessing Greek Textual Resources with Large Language Models and Conversational AI

Konstantinos Skianis, Ph.D.
https://y3nk0.github.io

Department of Computer Science & Engineering
University of Ioannina, Greece

Friday 3/11/2023

About me

Education:

- Ptichion @Informatics, Athens University of Economics and Business, Greece
- Master MVA @École Normale Superieure Paris-Saclay, Paris, France
- Ph.D. @École Polytechnique, Paris, France

Professional experience:

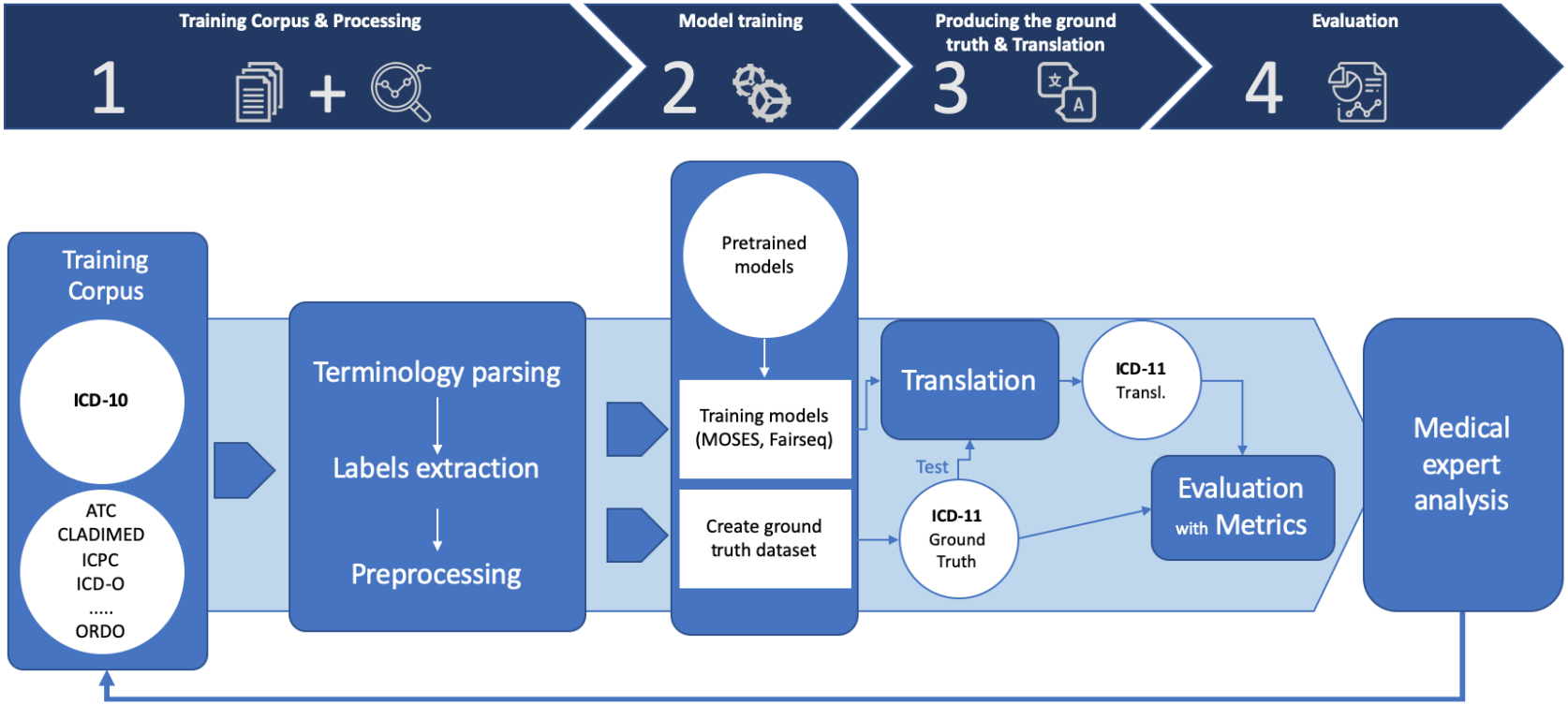
- Co-founder @BLUAI, PrimeHost
- Researcher @NKUA
- Adjunct Professor @AUEB
- Worked on multiple european projects

Highlights

- Best paper award at TextGraphs@NAACL 2018: "Fusing Document, Collection and Label Graph-based Representations with Word Embeddings for Text Classification"
- Nature Scientific Reports 2023 journal paper: "Predicting COVID-19 positivity and hospitalization with multi-scale graph neural networks"
- Both startups are members of well-known accelerators

BLUAI

Working for 3 years with the french Agence Numerique Sante in machine translation for medical terminologies



PrimeHost

Automate guest conversations with the power of AI

Greek document repositories

Have you tried searching information in diavgeia.gov.gr?

Motivation

Project

Build a pipeline which combines Large Language Models (LLMs) with conversational AI for Greek textual resources

Why?

- limited applications of cutting-edge NLP approaches in Greek resources
- most of them almost inacessible with old search capabilities

Goal

Facilitate extracting knowledge from Greek textual corpora:

- enabling asking questions
- assisting in creating/enhancing existing resources (knowledge graphs, ontologies etc.)
- models and methods can be used in multiple scenarios

Background

NLP has been greatly empowered by:

- Neural probabilistic language model (Bengio et al., 2000)
- Word embeddings (Mikolov et al., 2013)
- Transformers (Vaswani et al., 2017)
  - encoder-decoder architecture
  - encoder: multi-head self-attention layers to encode input for creating its latent representations
  - decoder: cross-attention on latent representations and autoregressively generates target
  - BERT: Pre-training of deep bidirectional transformers for language understanding (3TB data, 300M params) (Devlin et al., 2018)
- Large Language Model (LLM)
  - OpenAI GPT: Improving language understanding by generative pre-training (GPT3: 45TB data, 175B params) (Radford et al., 2018)

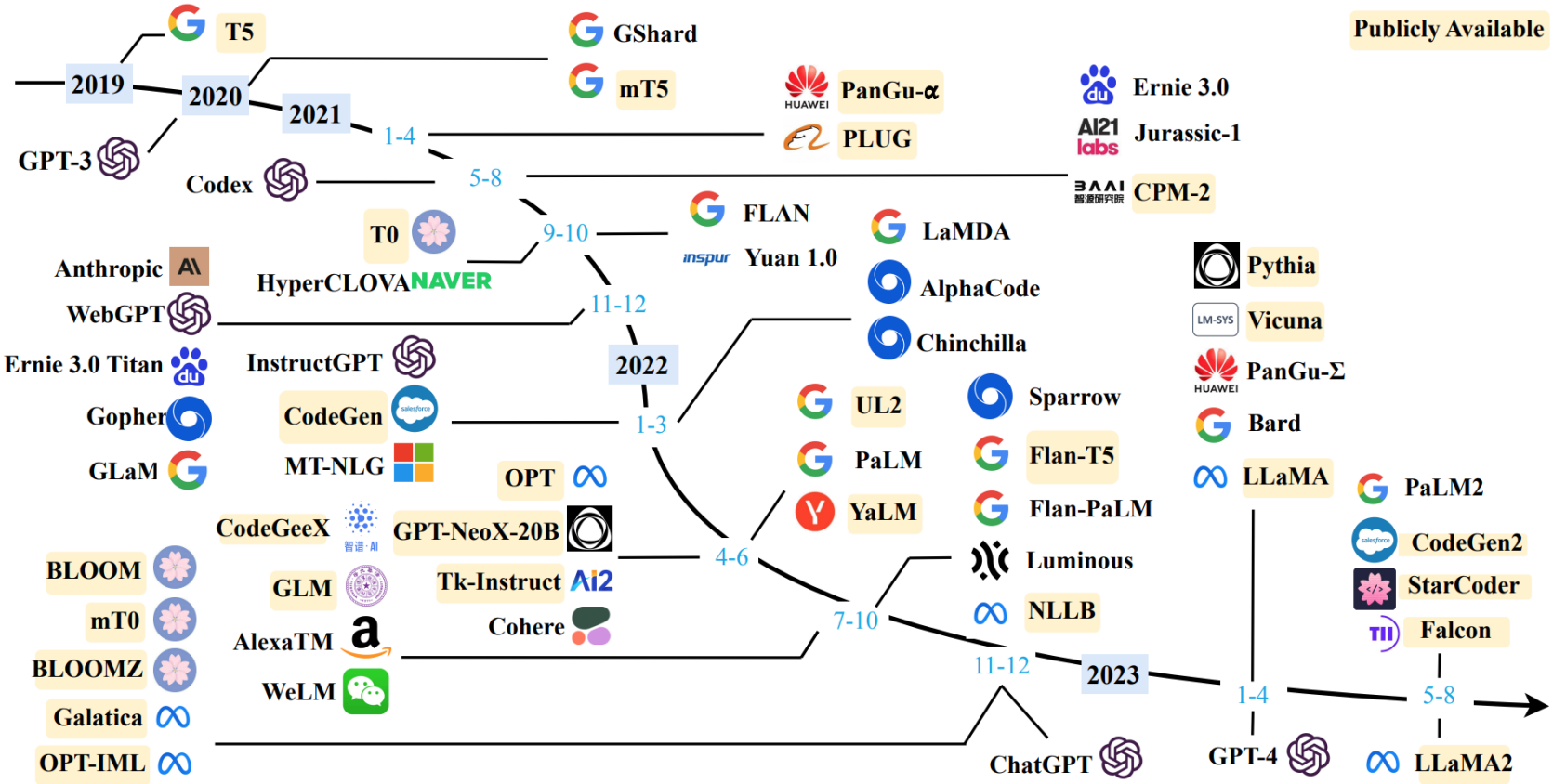
Foundational models

- Play a crucial role in forming the basis for more intricate AI and ML models and applications
- LLMs (e.g. BERT, GPT4) are prime examples of foundational models, that acquire knowledge from datasets, allowing them to capture fundamental patterns and relationships
- Initially pre-trained on extensive, general-purpose data and later adapted to specific tasks or domains
- Most commonly used objective to pre-train decoder-only LLMs, e.g. GPT3 and PaLM (Chowdhery et al., 2022), the language modeling task (LM)
- Given a sequence of tokens $x = \{x_1, \ldots, x_n\}$, the LM task autoregressively predicts the target tokens x_i based on the preceding tokens $x_{<i}$ in a sequence, maximize:

$$\mathcal{L}_{LM}(x) = \sum_{i=1}^{n} \log P(x_i | x_{<i}) \quad (1)$$

Fine-tuning and scale

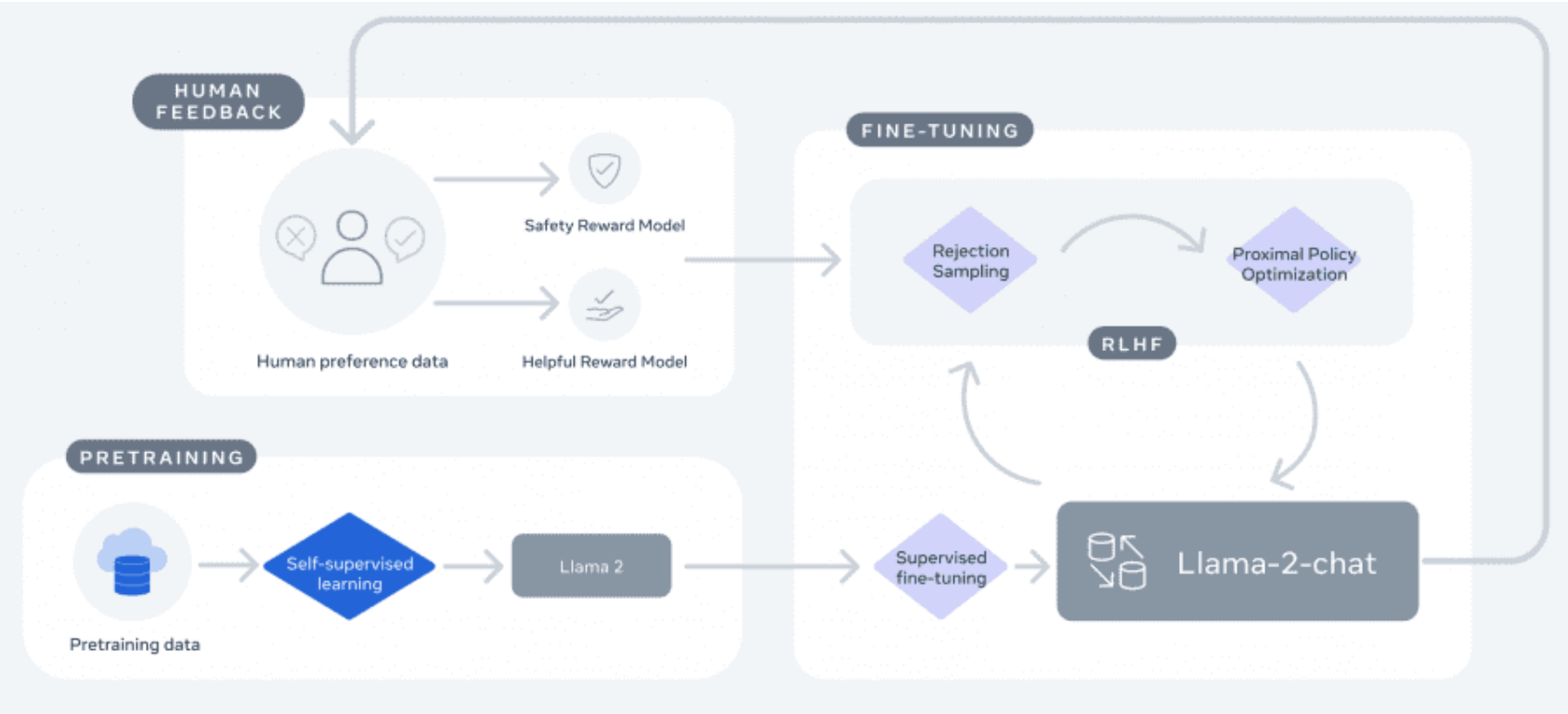
- Causal decoder architecture (GPT3) can achieve a superior zero-shot and few-shot generalization capacity
- Full fine-tuning $\longrightarrow$ retraining all model parameters is less feasible (e.g. deploying fine-tuned GPT-3 with 175B params is prohibitively expensive)



Timeline of existing LLMs (having a size larger than 10B) (Zhao et al., 2023)

LLaMA 2

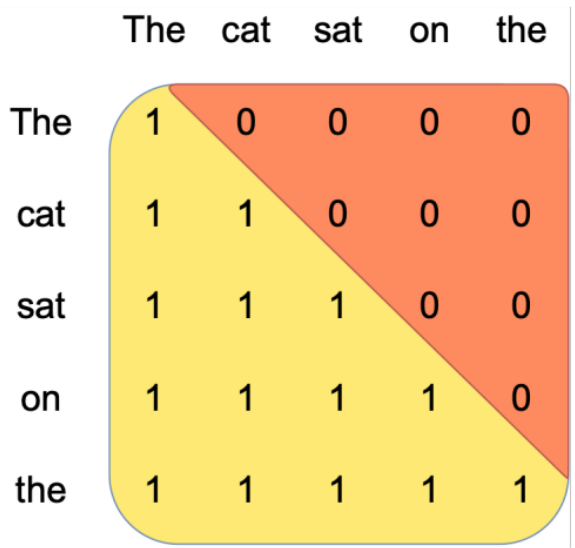
- Developed by META, open-source (Touvron et al., 2023)
- 70B parameters, 2T tokens, 2000 80G A100
- Grouped-query attention, Ghost Attention, In-Context Temperature re-scaling and Temporal Perception



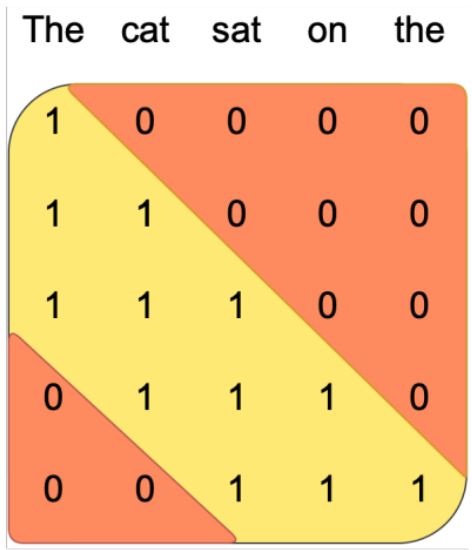
LLaMA 2-CHAT via supervised fine-tuning, Reinforcement Learning with Human Feedback (RLHF)

Mistral 7B

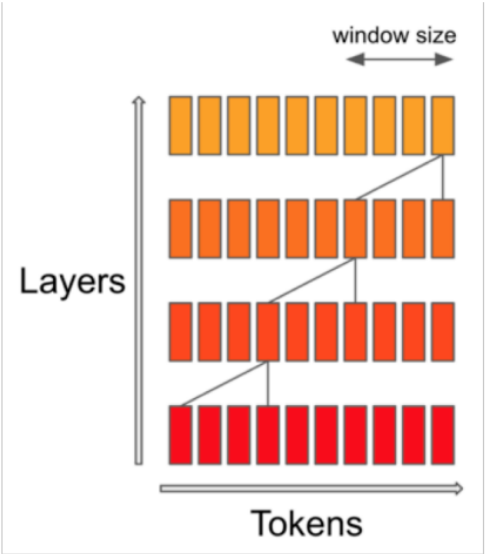
- Transformer-based, 7B params (Jiang et al., 2023)
- Outperforms Llama 2 13B on all benchmarks, Llama 1 34B on many, approaches CodeLlama 7B performance on code
- Grouped-query attention, Sliding Window Attention (SWA) to handle longer sequences at smaller cost
- SWA exploits the stacked layers of a transformer to attend in the past beyond the window size
- Higher layers have access to information further in the past than what the attention patterns seems to entail



Vanilla Attention



Sliding Window Attention



Effective Context Length

Large Greek textual resources

- A Greek Parliament Proceedings Dataset for Computational Linguistics and Political Analysis (Dritsa et al., 2022)
  - 1989 up to 2020
  - >1M speeches with extensive meta-data
  - extracted from 5,355 parliamentary sitting record files
- Corpus of Greek legislation, published by the National Publication Office,
- Corpus of EU legislation (Greek translation), as published in Eur-Lex
- Diavgeia (`diavgeia.gov.gr`):
  - 58M documents
  - 5260 active organizations/entities
  - 137K users
  - hard to access in a meaningful way

Related work

Models

- Greek-BERT (Koutsikakis et al., 2020)
  - 340M params, 29GB, 3.04B tokens
- Greek-BART: denoising auto-encoder (Evdaimon et al., 2023)
  - 181M params, 77GB

Systems

- A linked data platform for Greek legislation (Angelidis et al., 2018)
- AI-guided chatbots for citizens-government communication (Androutsopoulou et al., 2019)
- Chatbots to provide public service information (Stamatis et al., 2020)
- Chatbot for passport info (Antoniadis and Tambouris, 2021)
- Extracting Structured Information From Diavgeia (Tsironidou, 2021)

No models created by document repos and limited chatbot capabilities

Proposed work

Goal

Incorporate LLMs and conversational AI to exploit Greek textual resources

Case study

Ask questions to Diavgeia in natural language and extract knowledge

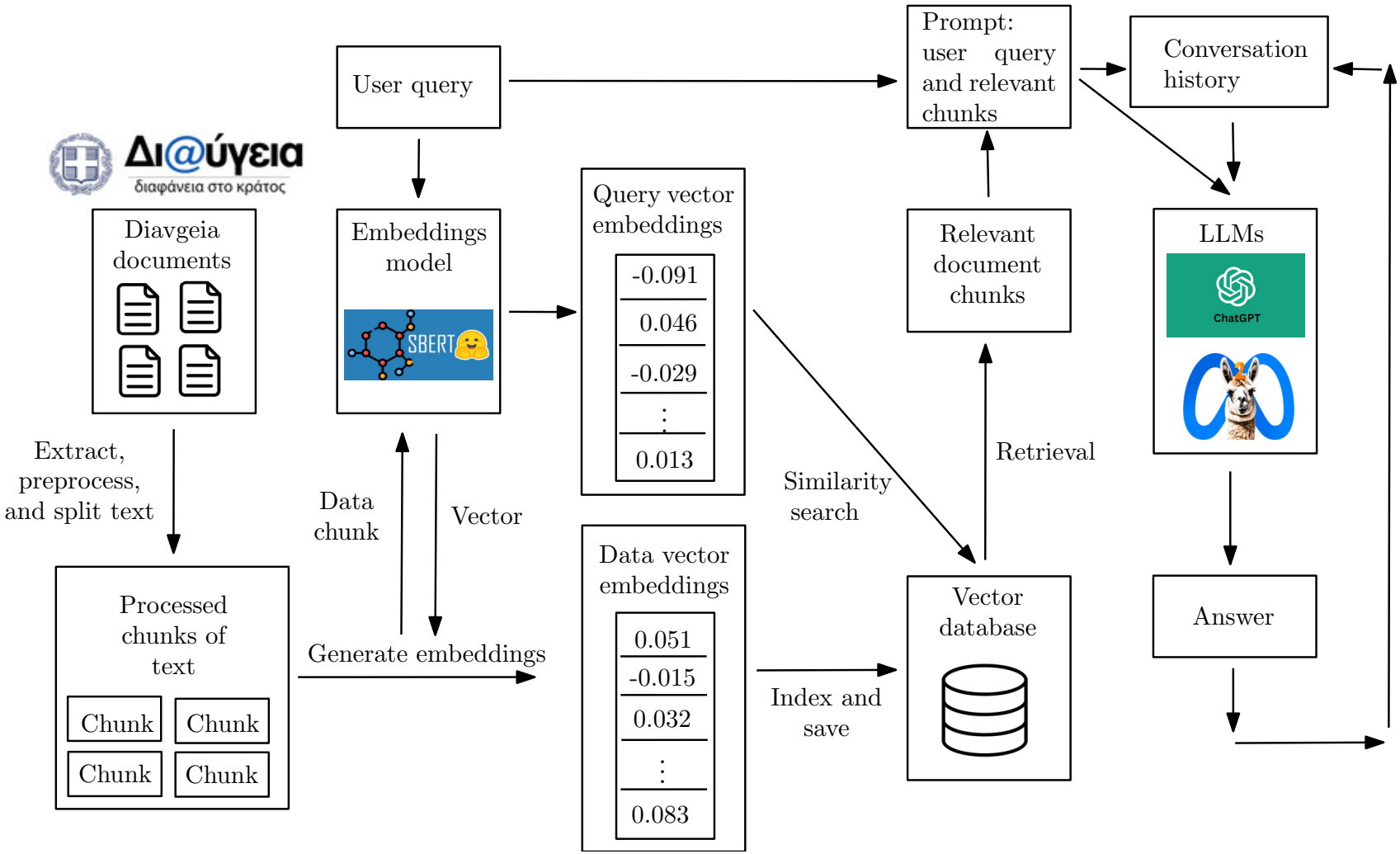
Proposed pipeline:

(1) download, preprocess and store documents,

(2) build a Retrieval-Augmented Generation (RAG) system, incorporating external knowledge into LLMs (Lewis et al., 2020)

(3) use LLM (e.g. ChatGPT API) and get answers

Pipeline



A full schematic illustration, based on: https://ubuntu.com/blog/llms-explained

Retrieval-Augmented Generation

Benefits

- **Consistency:** more likely to get the same answer from the same question
- **Increased factual accuracy:** LLM's responses are based on the provided information $\rightarrow$ model is less likely to hallucinate or mislead
- **Decreased cost:** building a RAG pipeline is less expensive than fine-tuning. Update only the database instead of training a new model
- **Currency:** ensure the LLM's responses are based on up-to-date data
- **Easily verifiable source:** access to the source for cross-checking
- **Strong scalability:** millions of documents

Cons

- **Complexity:** chunking, quality of search, low-quality data
- **Context length limitation:** ChatGPT (gpt-3.5-turbo) has a maximum context length of 4096 tokens
- **Limited creativity:** losing the point of generative AI

Implementation

Setup:

- 100 documents (1/1/2023 to 2/1/2023)
- BM25 retriever

Created a docker image with 4 services:

- Redis: open-source, key-value in-memory database to hold history
- Elasticsearch: search engine to look into documents
- Kibana: visualization, data analytics
- Flask API: RESTful API to expose endpoints
  - Response uses OpenAI API (gpt-3.5-turbo)

Example

Url: https://diavgeia.gov.gr/decision/view/93AB469HΓΥ-7IZ

ΑΔΑ	**93ΑΒ469ΗΓΥ-7ΙΖ**
Κατάσταση	Αναρτηθείσα
Ημερομηνία ανάρτησης	12/10/2023 14:51:02
Τελευταία τροποποίηση	12/10/2023 14:51:02
Ηλεκτρονικό αρχείο	Λήψη αρχείου | Προβολή αρχείου
Είδος	ΕΓΚΡΙΣΗ ΔΑΠΑΝΗΣ
Θέμα	ΠΡΟΪΟΝΤΙΚΕΣ ΔΩΡΕΕΣ
Θεματικές κατηγορίες	• ΟΙΚΟΝΟΜΙΚΕΣ ΚΑΙ ΕΜΠΟΡΙΚΕΣ ΣΥΝΑΛΛΑΓΕΣ
Αρ. πρωτοκόλλου	ΑΝΚ1374
Ημερομηνία έκδοσης	01/01/2023
Φορέας	ΕΘΝΙΚΟ ΘΕΑΤΡΟ
Οργανωτικές Μονάδες	• ΚΑΛΛΙΤΕΧΝΙΚΗ ΔΙΕΥΘΥΝΣΗ
Υπογράφοντες	• ΜΟΣΧΟΣ ΙΩΑΝΝΗΣ

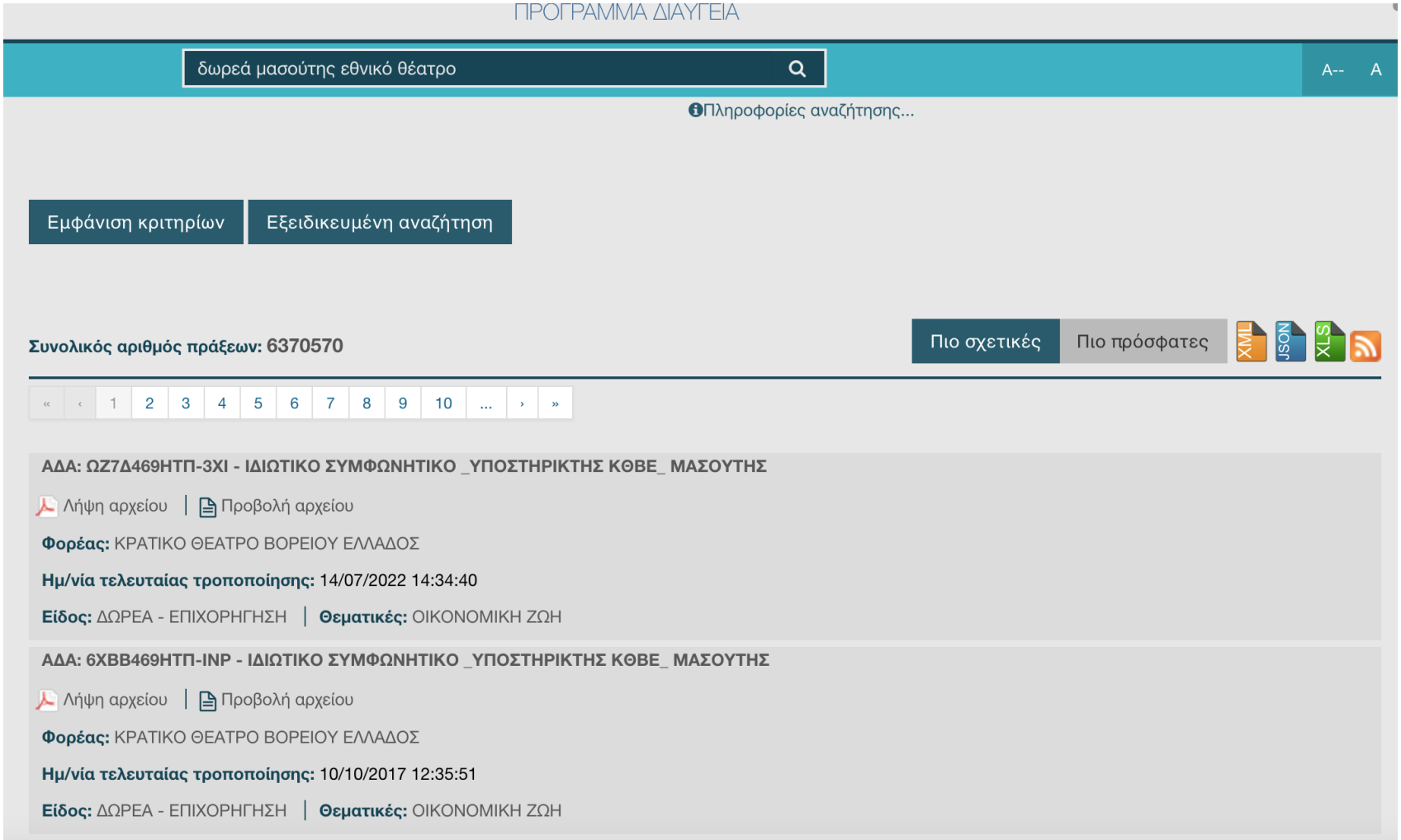
ΑΦΜ και επωνυμία φορέα

ΑΦΜ	090025586
Τύπος ΑΦΜ	Εθνικό
Επωνυμία φορέα	ΕΘΝΙΚΟ ΘΕΑΤΡΟ ΝΠΙΔ
Κράτος -μέλος της ΕΕ	
Οργανισμός χωρίς ΑΦΜ	

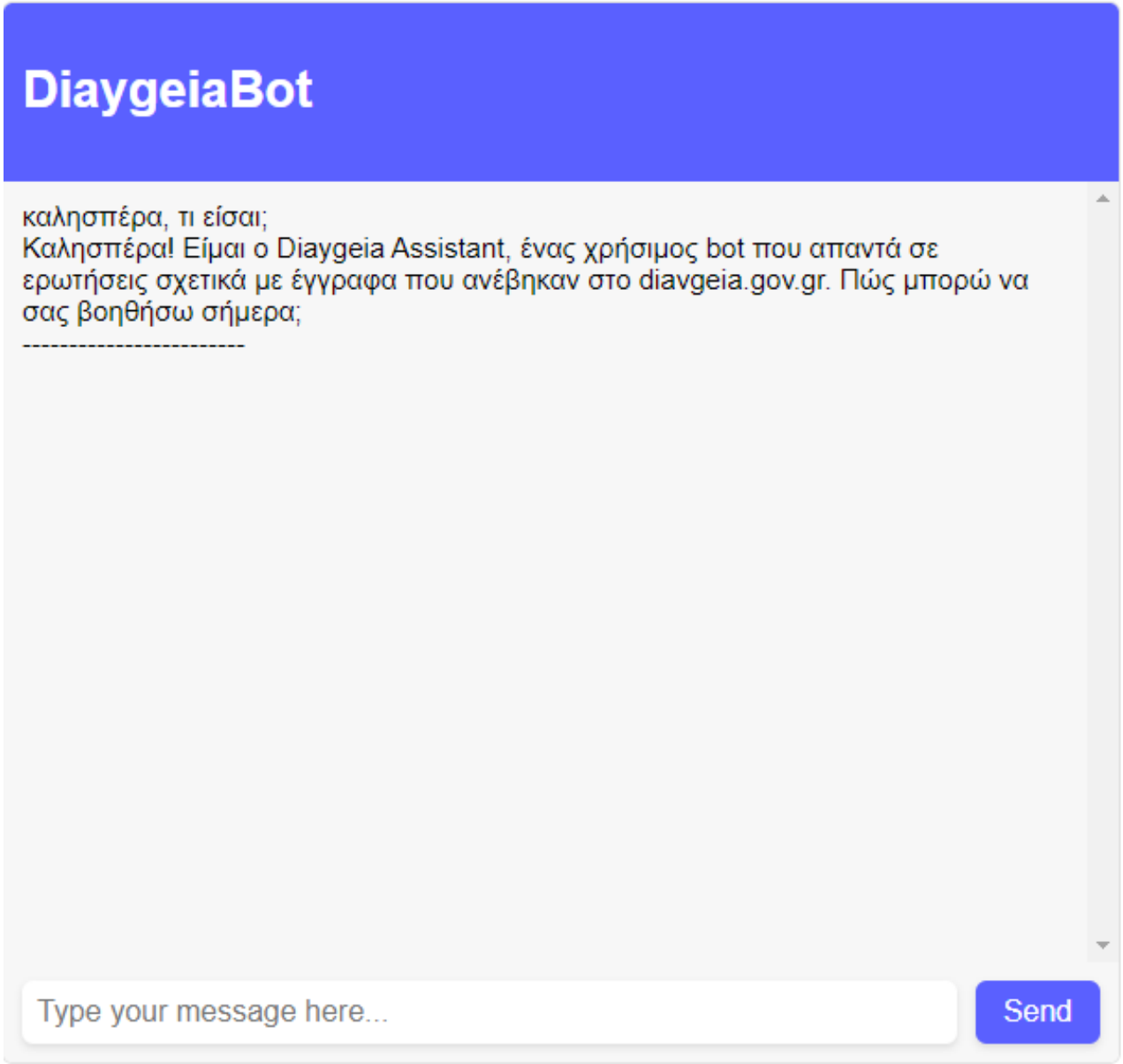
Στοιχεία αναδόχων

ΑΦΜ / Επωνυμία					**Ποσό δαπάνης**		**CPV**	**Αριθμός ΚΑΕ/ΑΛΕ**
ΑΦΜ	**Τύπος ΑΦΜ**	**Κράτος -μέλος της ΕΕ**	**Επωνυμία**	**Οργανισμός χωρίς ΑΦΜ**	**Αξία**	**Νόμισμα**		
					60.000,00	Ευρώ		

Search in diavgeia.gov.gr

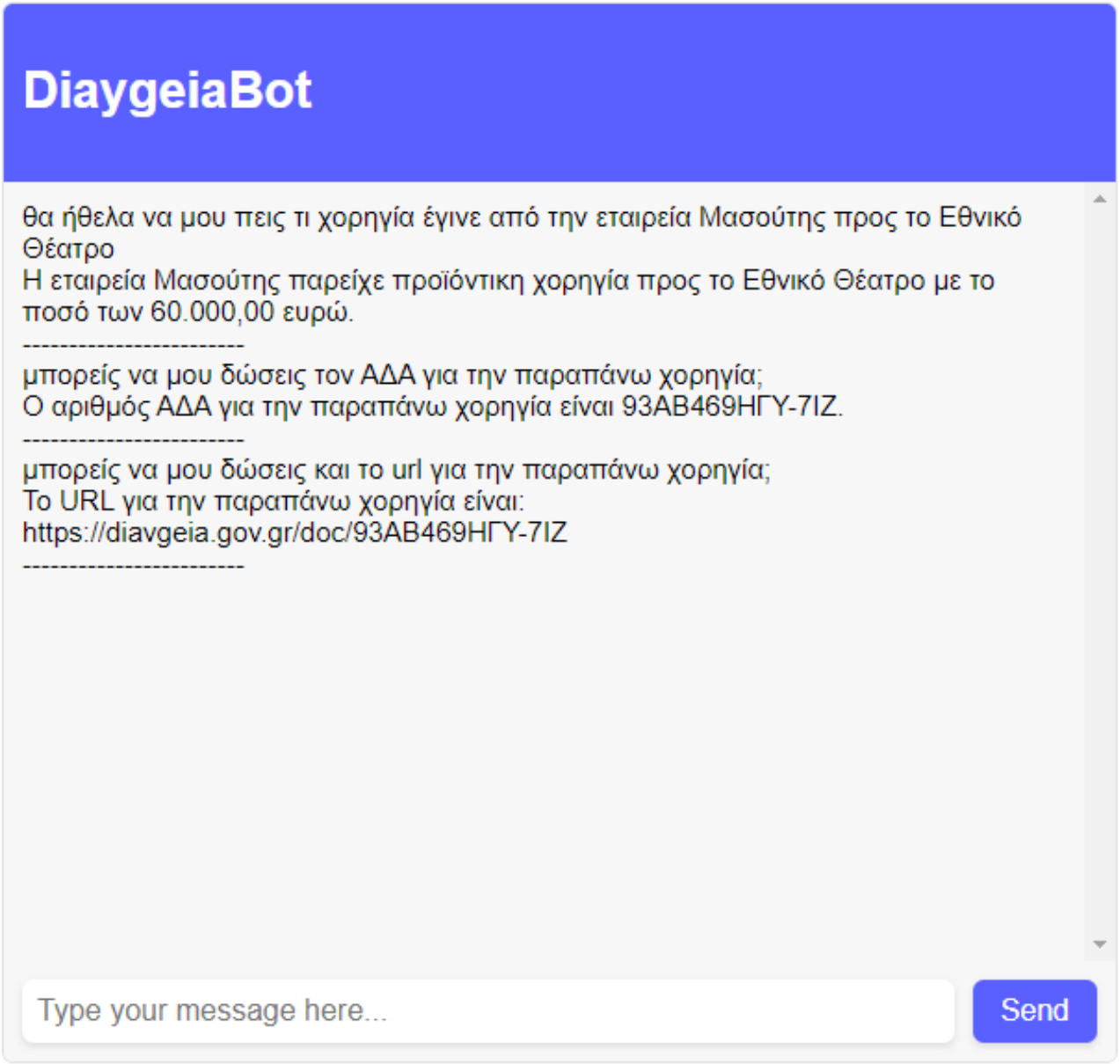


Frontend


DiaygeiaBot
καλησπέρα, τι είσαι;
Καλησπέρα! Είμαι ο Diaygeia Assistant, ένας χρήσιμος bot που απαντά σε ερωτήσεις σχετικά με έγγραφα που ανέβηκαν στο diavgeia.gov.gr. Πώς μπορώ να σας βοηθήσω σήμερα;

Type your message here...
Send

Extracting info via the chatbot



Key takeaways

Diavgeia:

- search returns thousands of docs
- not the most relative results

LLM-based conversational AI with RAG:

- Prompt engineering can be crucial
- Hallucination (e.g. showing doc with ADA that does not exist)
- Context size affects performance and cost
- Conversation history plays important role
- GPT4 more powerful, but expensive

To-do:

- More testing, embeddings, improve UI
- Evaluation: strong LLMs as judges to evaluate on more open-ended questions (Zheng et al., 2023)

Next improvements: fine-tuning

Reminder

Full fine-tuning LLMs remains expensive and slow.

Parameter-efficient fine-tuning (PEFT), cost-efficient, faster, adapting pre-trained models to downstream applications, by fine-tuning a small number of extra model parameters:

- Parameter-efficient prompt tuning (Lester et al., 2021)
- LoRA: Low-rank adaptation of large language models (Hu et al., 2022);(Dettmers et al., 2023)
- P-tuning v2: Prompt tuning comparable to fine-tuning across scales and tasks (Liu et al., 2022)
- GPT understands, too (Liu et al., 2023)

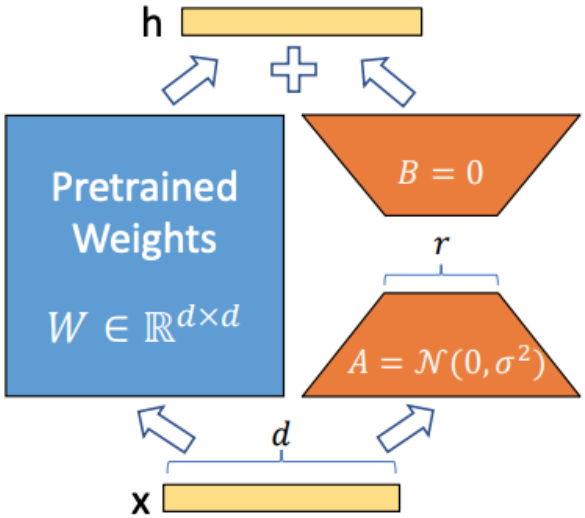


Figure: LoRA reparametrization. Only A and B are trained.

Summary

Contribution:

- first approach to use an end-to-end NLP approach, combining LLMs, conversational AI and Greek document repos
- real use-case with Diavgeia, access it better

Why is it important?

- more resources can be added or used for finetuning
- no training required (e.g. with OpenAI API)
- free open-source alternative (e.g. LLaMa 2, Mistral 7B)
- the pipeline can be easily adopted for other documents repos
- highly adjustable for other tasks (e.g. enhance corpora, create new legal documents from scratch fast)

Future directions?

- Regularization
  - Opt-iml: Scaling language model instruction meta learning through the lens of generalization (Iyer et al., 2022)
- Knowledge graphs
  - Temporal Question Answering over KGs (Chen et al., 2023)
  - Graph-Aware Language Model Pre-Training (Xie et al., 2023)
- Graph Neural Networks
  - Graph Neural Prompting with LLMs (Tian et al., 2023)
- Multimodality
  - Palm-e: An embodied multimodal LLM (Driess et al., 2023)
  - Adapted Multimodal Bert with Layer-Wise Fusion for Sentiment Analysis (Chlapanis et al., 2023)

Thank you! Questions?

References I


Androutsopoulou, Aggeliki et al. (2019). "Transforming the communication between citizens and government through AI-guided chatbots". In: *Government information quarterly* 36.2, pp. 358–367.


Angelidis, Iosif et al. (2018). "Nomothesia: a linked data platform for Greek legislation". In:


Antoniadis, Panteleimon and Efthimios Tambouris (2021). "PassBot: A chatbot for providing information on Getting a Greek Passport". In: *Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance*, pp. 292–297.


Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent (2000). "A neural probabilistic language model". In: *Advances in neural information processing systems* 13.


Chen, Ziyang, Jinzhi Liao, and Xiang Zhao (2023). "Multi-granularity Temporal Question Answering over Knowledge Graphs". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11378–11392.


Chlapanis, Odysseas S, Georgios Paraskevopoulos, and Alexandros Potamianos (2023). "Adapted Multimodal Bert with Layer-Wise Fusion for Sentiment Analysis". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.

References II

Chowdhery, Aakanksha et al. (2022). "Palm: Scaling language modeling with pathways". In: *arXiv preprint arXiv:2204.02311*.

Dettmers, Tim et al. (2023). "Qlora: Efficient finetuning of quantized llms". In: *arXiv preprint arXiv:2305.14314*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Driess, Danny et al. (2023). "Palm-e: An embodied multimodal language model". In: *arXiv preprint arXiv:2303.03378*.

Dritsa, Konstantina et al. (2022). "A Greek Parliament Proceedings Dataset for Computational Linguistics and Political Analysis". In: *Advances in Neural Information Processing Systems* 35, pp. 28874–28888.

Evdaimon, Iakovos et al. (2023). "GreekBART: The First Pretrained Greek Sequence-to-Sequence Model". In: *arXiv preprint arXiv:2304.00869*.

Hu, Edward J et al. (2022). "Lora: Low-rank adaptation of large language models". In: *ICLR*.

Iyer, Srinivasan et al. (2022). "Opt-iml: Scaling language model instruction meta learning through the lens of generalization". In: *arXiv preprint arXiv:2212.12017*.

Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL].

References III

Koutsikakis, John et al. (2020). "Greek-bert: The greeks visiting sesame street". In: *11th Hellenic conference on artificial intelligence*, pp. 110–117.

Lester, Brian, Rami Al-Rfou, and Noah Constant (2021). "The power of scale for parameter-efficient prompt tuning". In: *EMNLP*.

Lewis, Patrick et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33, pp. 9459–9474.

Liu, Xiao et al. (2022). "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks". In: *ACL*.

Liu, Xiao et al. (2023). "GPT understands, too". In: *AI Open*.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26.

Radford, Alec et al. (2018). *Improving language understanding by generative pre-training*. OpenAI.

Stamatis, Anestis et al. (2020). "Using chatbots and life events to provide public service information". In: *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, pp. 54–61.

Tian, Yijun et al. (2023). "Graph Neural Prompting with Large Language Models". In: *arXiv preprint arXiv:2309.15427*.

References IV

- Touvron, Hugo et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL].
- Tsironidou, Eleni (2021). *Extracting Structured Information From Diavgeia Portal*.
- Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.
- Xie, Han et al. (2023). "Graph-Aware Language Model Pre-Training on a Large Graph Corpus Can Help Multiple Graph Applications". In: *KDD*.
- Zhao, Wayne Xin et al. (2023). "A survey of large language models". In: *arXiv preprint arXiv:2303.18223*.
- Zheng, Lianmin et al. (2023). "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena". In: *NeurIPS*.