Towards an OSS LLM for Greek

Institute for Language and Speech Processing

28/6/2023

ATHENA Research & Innovation Information Technologies

What are Language Models (LMs) and Large LMs (LLMs)

- Language models are deep neural networks that specialize in predicting the next word within a sequence of text
- Scaling LMs (model and training data size) lead to LLMs
  - Use the Transformer architecture
  - Have improved model capacity on various tasks
  - Can solve few-shot tasks through in-context learning

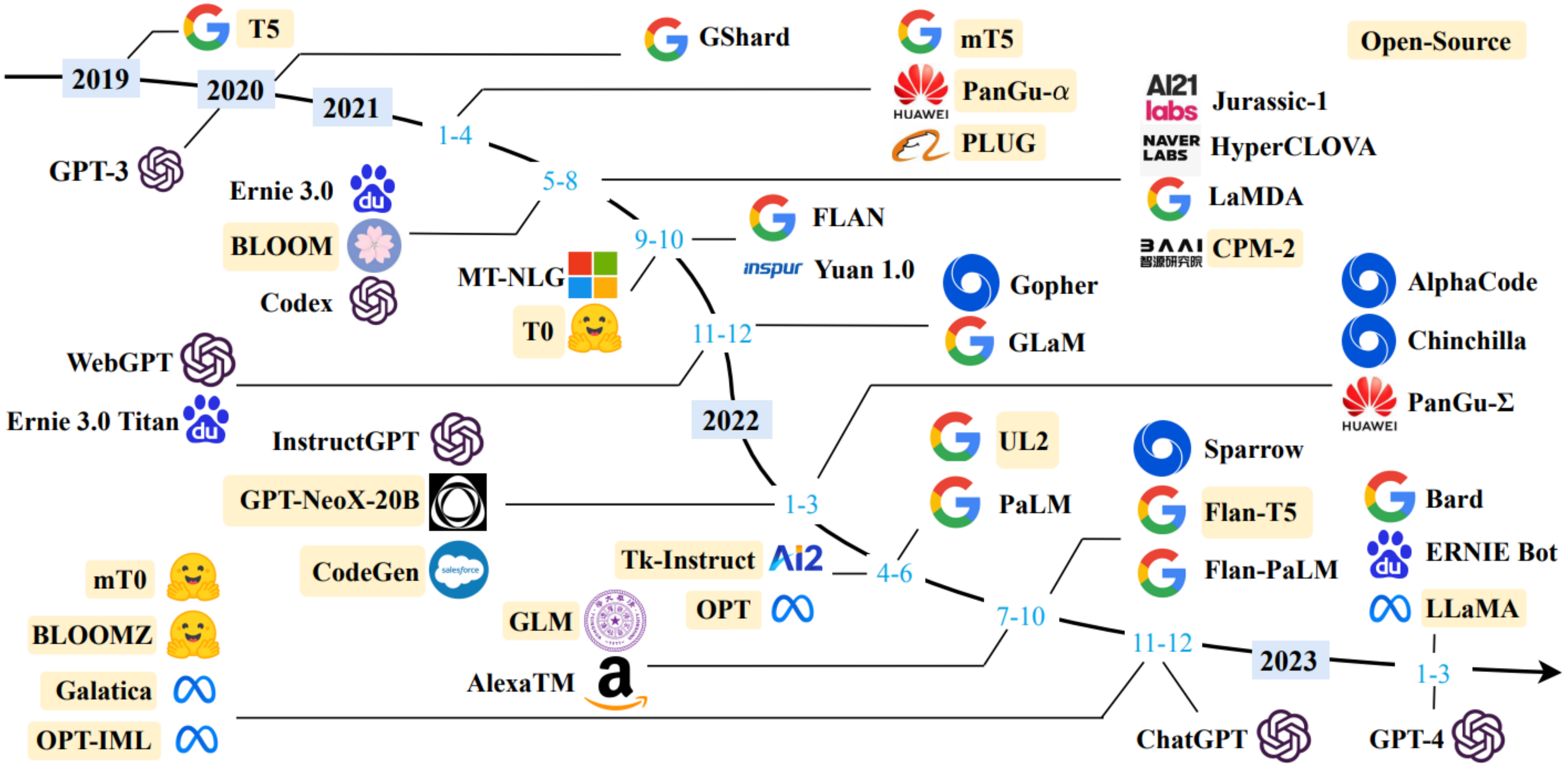

Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

Source: A Survey of Large Language Models, 2023, WX Zhao et al. (https://arxiv.org/abs/2303.18223)

Motivation for an OSS LLM

- Data Privacy and Security
- Dependency and Customization
- Cost and Scalability
- Access and Availability

The Greek language in existing multilingual LLMs

- There are a few multilingual LLMs available, but most use just a small portion of Greek data (BLOOM, Pythia) or even no Greek data at all (LLaMA)
- Models like GPT-4 seem to perform well on Greek data, but are not open

General steps towards an OSS LLM for Greek

- Data collection and preprocessing
- Pre-training (if building an LLM from nothing)
- Fine-tuning / in-context learning
- Evaluation (over a variety of tasks)

Data collection and preprocessing

- Collect available open Greek datasets from various sources
  - Avoid bias by including a diverse range of datasets
  - Licenses must be taken into consideration
- Clean and pre-process the collected data
  - Remove non-Greek content, noise, duplicates
  - Ensure that our dataset is of a high-quality

Pre-training

- Using the collected Greek text, we train a language model using one of the available techniques
- Requires huge computational resources for training and inference
- Deployment is very challenging
  - Memory constraints
  - Optimization

Fine-tuning

- Process of further training a pre-trained model on new data to
  - Improve its performance on one specific task
  - Adapt the model in a new domain
- One approach would be to
  - Select one (or more) of the available multilingual LLMs (from the Hugging Face library)
  - Fine-tune the model on a high-quality Greek corpus
  - That model could then be further adapted to a wide range of downstream tasks
- The new model would also have the knowledge that the pre-trained model has
- By fine-tuning a model for a specific task, we need to keep a separate copy of the entire model for each of these tasks

multilingual LLMs that can be pre-trained for "cheap" and contain Greek data

- **mBERT**: Only contains data from Greek Wikipedia
- **XLM-R (XLM-RoBERTa)**: Uses data from Common Crawl (47Gb) – now outperformed by **XLM-V**
- **mT5**: Uses the C4 corpus (a cleaner version of Common Crawl)

In-context learning through prompt creation

- Popularized in the original GPT-3 paper as a way for LMs to learn tasks with only a few examples
- Demonstrations of the task are provided to the model in natural language without any additional training
- The LM is given a prompt that consists of a list of input-output pairs demonstrating a specific task
- At the end of the prompt, a test input is added to allow the LM to predict the next tokens by conditioning on the prompt

When should in-context learning be used

- Useful when we have a small dataset, or only a few training examples
- Very helpful approach for tasks that require adaptability or when the model needs to generalize from a few examples
- Would it be sufficient for adapting a pre-trained LLM to Greek?

Fine-tuning vs. in-context learning

- Using prompts alone might result in less accurate outputs
- The base model might struggle with applying a task to Greek
- Fine-tuning results in better performance and higher accuracy

Evaluation of an LLM

- To achieve a comprehensive evaluation of an LLM we need to employ a set of benchmark tasks covering a wide range of language-related challenges
  - Question answering
  - Summarization
  - Text completion
  - Sentiment analysis
  - Machine translation etc.
- Example: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Evaluation of a Greek LLM

- Currently there are no evaluation datasets for Greek
- Options:
  - Try to build evaluation datasets for specific tasks using existing resources
  - Use machine translation to translate an established evaluation dataset
  - Use some general benchmark in multi-lingual NLP like Wino-X

Example of building a LM

Pretraining of GR-Electra PLM

Task description

- Train a Greek language model to use as a text encoder in a multimodal architecture
- Presented in Paraskevopoulos G, Pistofidis P, Banoutsos G, Georgiou E, Katsouros V. Multimodal Classification of Safety-Report Observations. Applied Sciences. 2022; 12(12):5781. https://doi.org/10.3390/app12125781

What is Electra

- A method for self-supervised language representation learning
- Used to pre-train transformer networks using little computing resources
- At small scale, Electra achieves strong results even when trained on a single GPU
- At large scale, Electra achieves state-of-the-art results on the SQuAD 2.0 dataset

Datasets Used

- Google's C4 (Colossal Clean Crawled Corpus) dataset
  - 190 Gb is the uncompressed text size of Greek text
  - Described as **clean-ish data**
- The Greek version of Wikipedia
- The Hellenic National Corpus
  - Developed by ILSP
  - Consists of written texts with almost 97.000.000 words

Corpus preparation

- Filering steps:
  - Lines that contain URLs
  - Lines that do not contain Greek characters
  - Lines that contain UTF-8 characters that do not belong in one of the following sets: (Greek, Latin, Numeric, Punc., accents)
  - Lines with Ancient or purist Greek text, by filtering diacritics
  - Use of unigram or bigram LM to break words that are merged during crawling (e.g. "somewordsaremerged")

Model training

- Normalization: Remove accents and convert to lowercase (electra-base-uncased)
- Used just a 20GB subset of the corpus
- AdamW optimizer
- Learning Rate: 10^{-4}
- Warmup: 8000 steps
- Batch size: 16
- Trained for 18 days on 4 NVIDIA RTX 3090 GPUs

Έρευνα & Καινοτομία
Τεχνολογίες Πληροφορίας

ΑΘΗΝΑ

ATHENA

Research & Innovation
Information Technologies