Training the Greek LLM Meltemi

Vassilis Katsouros
Director
Institute for Language and Speech Processing
ATHENA Research & Innovation Information Technologies

EuroCC@Greece – HPC Training Series – January 17, 2025

Outline

- Introduction
- Data
- Training procedure foundation and chat models
- Evaluation
- Model deployment
- Next Steps

Introduction

- Openly available LLMs do not focus on languages with unique scripts, like Greek, although they have seen Greek text data.
- Training an LLM from scratch is complex and requires significant resources (compute, data, etc.).
- A more cost-effective approach is the continual pretraining of an existing foundation model.
- Towards this goal, we selected **Mistral-7B of Mistral AI** as our base model.

1 https://mistral.ai/news/announcing-mistral-7b/

Motivation and Expected Impact

1. **Democratization of AI Technology:** Open alternatives to commercial solutions help with the proliferation of AI technologies to everyone enabling the use of cutting-edge technologies without the prohibitive costs.
2. **Combating digital under-representation of languages:** Developing technologies for less popular languages, like Greek, helps to preserve cultural heritage and provide tools for education, communication, and content creation in them.
3. **Transparency and Trust:** Open-source models promote transparency in AI development, allowing the community to inspect, verify, and improve the models.
4. **Community-Driven Improvements:** The community can contribute to open-source models, leading to more robust, versatile, and domain-aware AI solutions.
5. **Economic and Educational Opportunities:** Open LLMs empower a wider audience to develop AI skills, fostering economic growth and providing educational opportunities.

Challenges

Aggregating large amount of high-quality dataset of Greek texts

- **Availability:** Finding a diverse and comprehensive collection of Greek texts.
- **Quality:** Ensuring the dataset is high-quality.
- **Licensing:** Copyright issues to use or share these texts.

1 https://mistral.ai/news/announcing-mistral-7b/

Challenges

Forming a team with versatile skills and expertise

- **Skill Diversity:** Team with skills in NLP, AI algorithms, Greek linguistics, data engineering, and software development.
- Collaboration: Ensuring effective collaboration among team members with different expertise and backgrounds.

1 https://mistral.ai/news/announcing-mistral-7b/

Challenges

Find the necessary computing resources

- **Cost:** High cost of computing resources required for training LLMs of many billion parameters.
- **Access:** Limited access to high-performance computing infrastructure.
- **Scalability:** Need for scalable solutions that can grow with the project's requirements.

1 https://mistral.ai/news/announcing-mistral-7b/

Selecting an appropriate name

Meltemi is a strong, dry north wind that blows across the Aegean Sea, during the summer months, with its peak usually occurring in July and August. Its intensity can vary from gentle breezes to strong gales, making it both a vital aspect of local weather and a significant factor in the region’s climate.

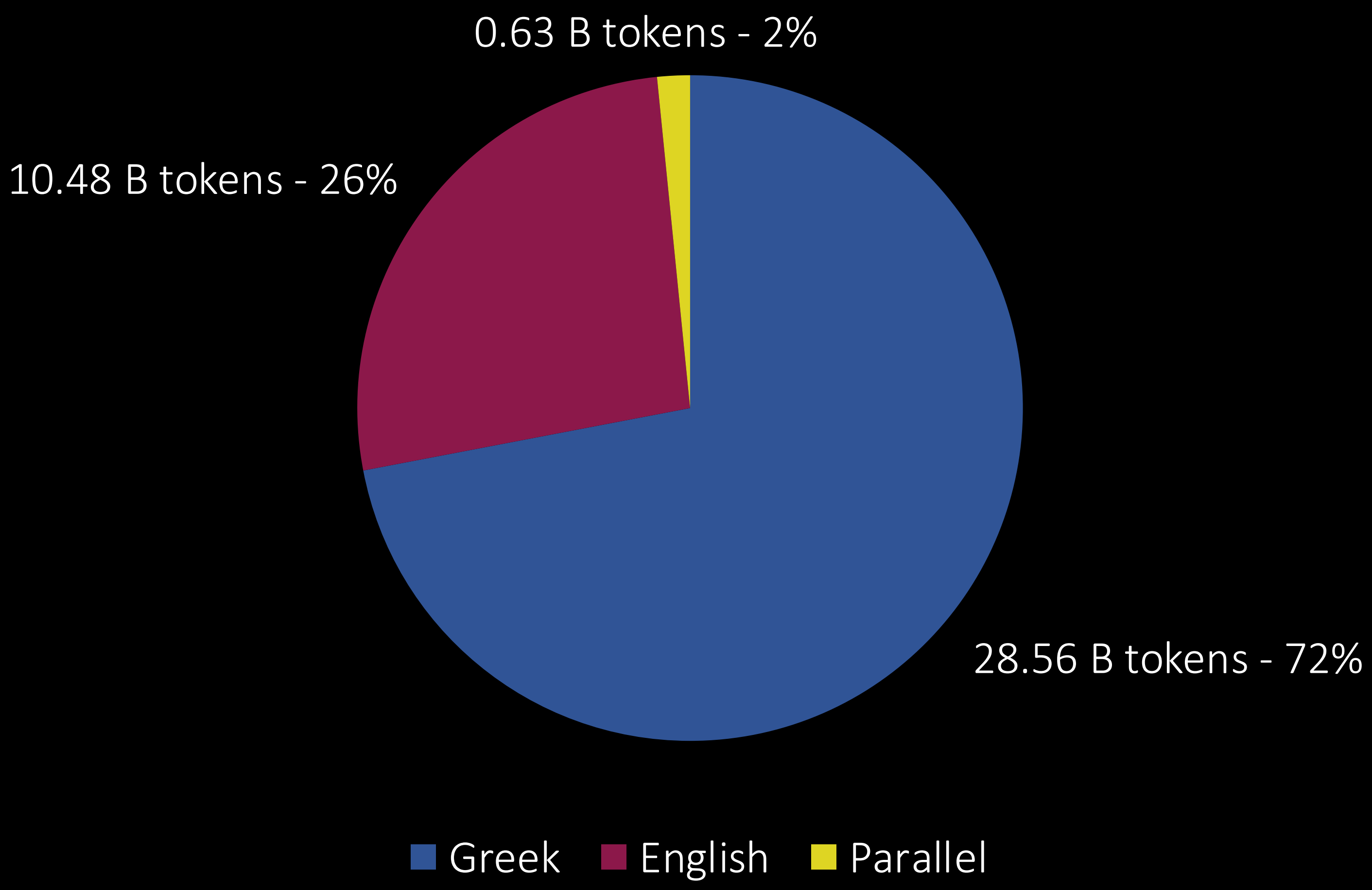
Training Dataset

Selection – Collection - Processing

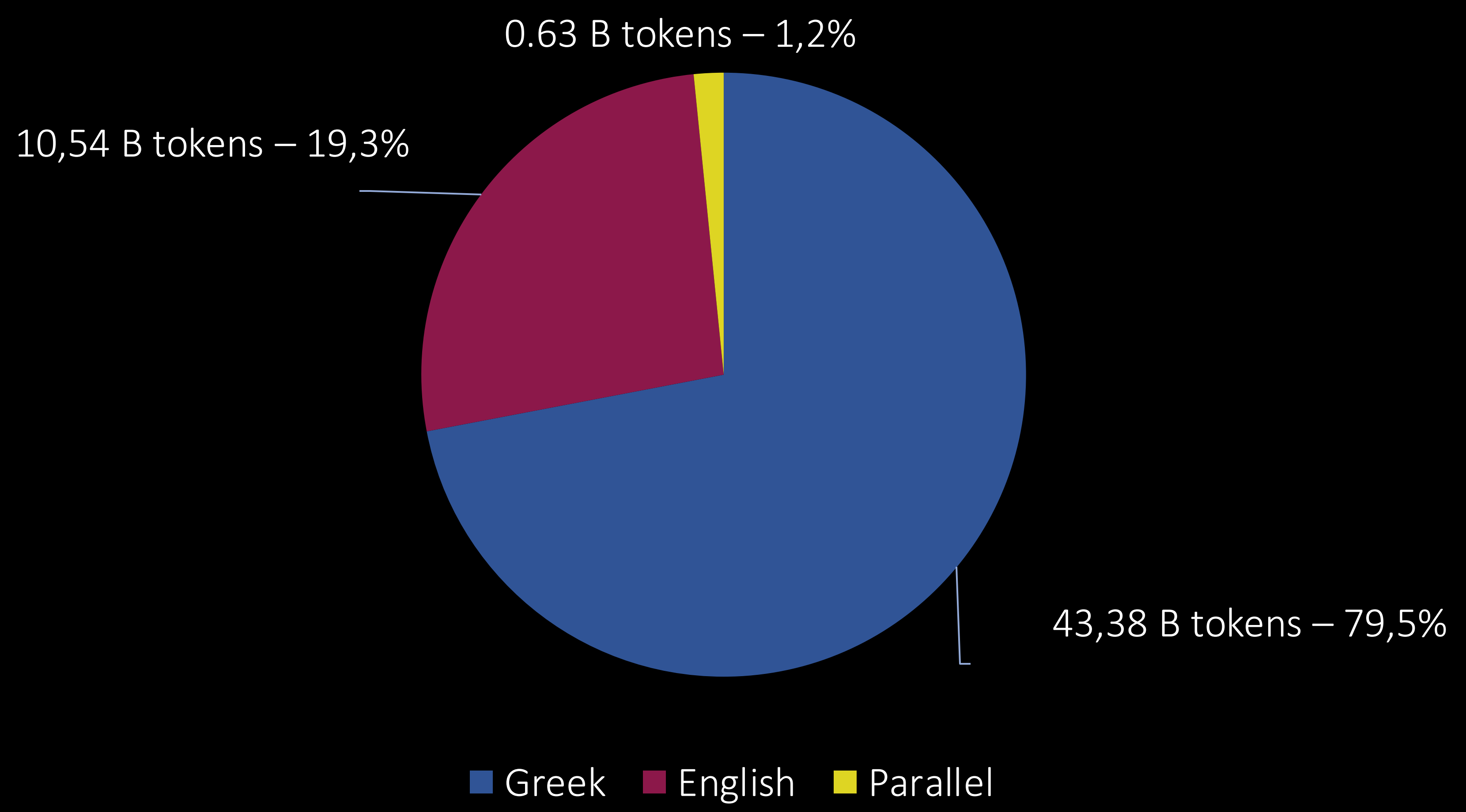
Dataset requirements for Meltemi

- LLMs require significant amounts of data for training
- Mistral 7B has seen a vast amount of data but not a lot of Greek
- For applying continual pretraining with Mistral 7B we need to:
  - collect as much as possible Greek text data of high-quality
  - add some English text data to tackle catastrophic forgetting
  - enrich dataset with parallel EN-EL data to:
    - learn the "relationship" between the two languages
    - be able to seamlessly switch between Greek and English in responses (if needed)

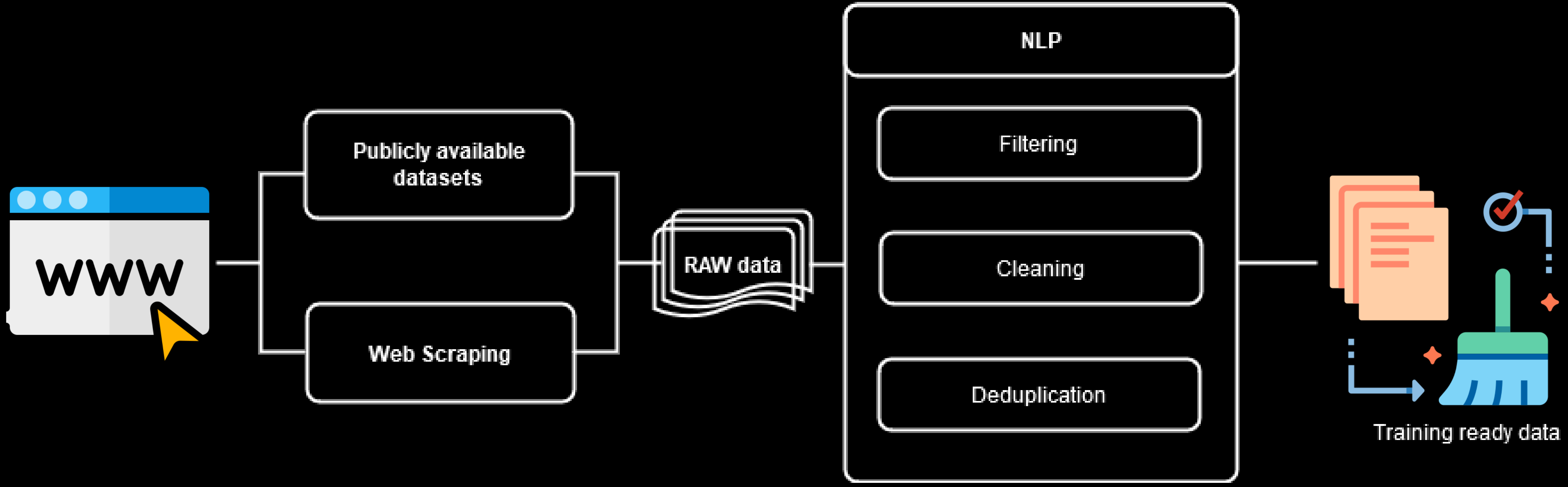
Composition of pretraining Data for v1



Composition of pretraining Data for v1.5



Data Preprocessing



Greek Data sources

- Collected high quality Greek monolingual texts from various publicly available data sources, including:
  - Wikipedia
  - ELRC-SHARE
  - Parliamentary proceedings
  - EUR-LEX
  - MaCoCu
  - CulturaX
  - Various academic repositories

Data Preprocessing

- Text extraction from PDFs & HTMLs, etc.
- Conversion in metadata-enriched JSON format
- Pre-processed & filtered using:
  - Rule-based filtering (e.g., min. word length, "lorem ipsum", etc.)
  - Scores & Thresholds, such as:
  - Fluency scores with KenLM models
  - Alignment scores for parallel data
- Document level deduplication (minhash LSH 5-grams)
- Ensured data distribution remains balanced throughout training

Training procedure

Steps

- Implemented a three-stage pretraining strategy that included
  - **Vocabulary extension**: Extend the Mistral tokenizer to include Greek tokens.
  - **Warm start embeddings**: Perform light fine-tuning step on the embeddings that correspond to the new tokens using 10% of the corpus. Other parameters are kept fixed.
  - **Continual pretraining**: Train all model parameters on the full training corpus.
- The training took 25 days
  - Consuming ~ 2,300 kWh
  - Including experimentation and failed runs (~8 days)
  - Gold run took ~17 days

Vocabulary Extension

- The original Mistral tokenizer did not contain meaningful Greek subwords
  - -> it performs character-level tokenization for Greek
  - Need to extend it for Greek subwords since this will limit the ability of the model to capture context and semantics

```
Text:
 Τα μεγάλα γλωσσικά μοντέλα χρειάζονται καλούς tokenizers


Tokenized with mistralai/Mistral-7B-v0.1:
['_', 'Τ', 'α', '_', 'μ', 'ε', 'γ', 'ά', 'λ', 'α', '_', 'γ', 'λ', 'ω', 'σ', 'σ', 'ι', 'κ', 'ά', '_', 'μ', 'ο', 'ν', 'τ', 'έ', '
λ', 'α', '_', 'χ', 'ρ', 'ε', 'ι', 'ά', 'ζ', 'ο', 'ν', 'τ', 'α', 'ι', '_', 'κ', 'α', 'λ', 'ο', 'ύ', 'ς', '_token', 'izers']


Tokenized with ilsp/Meltemi-7B-v1:
['_Τα', '_μεγάλα', '_γλωσσ', 'ικά', '_μοντέλα', '_χρειάζονται', '_καλούς', '_token', 'izers']
```

Vocabulary extension (cont.)

- Used a corpus containing 10M words
  - Stratified sampling across all the subcorpora
- Trained a sentencepiece model on this corpus
- Added new tokens to the tokenizer
  - Need to take care to not add double entries
  - If a token is already included we use the original one
- Original vocabulary size: 32000 subwords
- Extended vocabulary size: 61362 subwords

Vocabulary extension (cont.)

Tokenizer Model	Vocabulary Size	Fertility Greek	Fertility English
Mistral 7B	32.000	6,80	1,49
Meltemi 7B	61.362	1,52	1,44

Warm start embeddings

- The new tokens correspond to ~30k new **randomly initialized** rows in the embeddings matrix
- We need a better initialization strategy for the new embeddings to speed up training
  - Step 1: Calculate mean (μ) and variance (Σ) of the original embeddings
  - Step 2: Each new embedding vector is sampled from N(μ, Σ)
  - Step 3: Run a fine-tuning step for the new embeddings on 10% the corpus
    - All other parameters stay frozen

Continual pretraining vs Training from scratch

- CPT is an adaptation method where you continue to train on new data
  - Preserve knowledge from old data
  - Adapt to the new domain or language
  - Cheaper
  - Can lead to better performance than training from scratch due to transfer learning

Continual pretraining on Mistral-7B

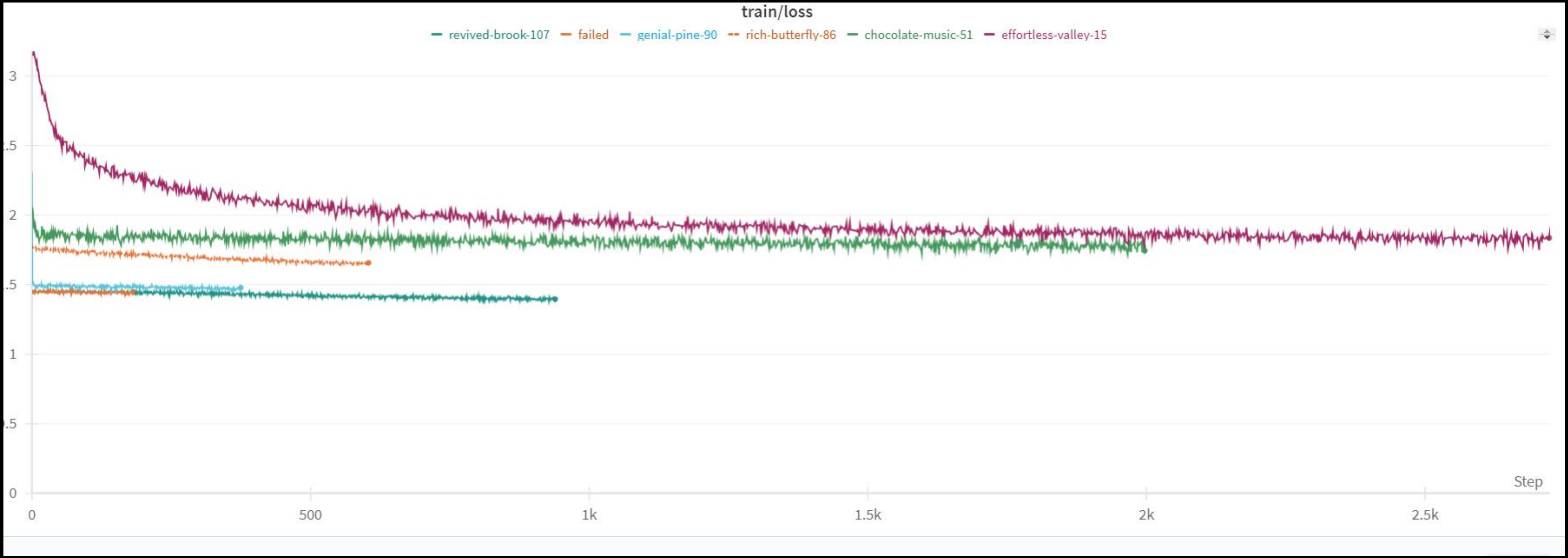
- Mistral-7B is a 7-billion parameter transformer
  - 32 layers
  - 4096 dimensions
  - 8192 context length
  - Sliding window attention, key-value caching, prefilling
- Officially very little is disclosed about the training data

- Why we chose it?
  - Good performance (at the time of creating Meltemi)
  - Apache2 license (open for research and commercial use)

Continual pretraining details

- Frameworks used:
  - Huggingface / torch -> Model and data
  - Deepspeed -> Multi-GPU training
- Batch size: 4.5M tokens
- Trained for 25000 steps

Training loss curves



Using an Amazon p5.48xlarge instance (8 x H100 GPUs)

Training logistics

- Working with LLMs and huge amount of data is expensive
- Hardware is not provided completely on-demand
  - They are reserved on a specific date
- Debugging is challenging
  - The code needs to be verified and working upon the reservation date
- Can’t base all decisions on experiments / ablations (bottom-up)
  - Most decisions are based on intuition and the literature (top-down)

Training outcome: Meltemi-7B-v1 / v1.5

- A foundation LLM for the Greek language that can be used for
  - Text Generation and Completion
  - Summarization
  - Translation
  - Question Answering
  - Text Classification
- The extent to which it performs these tasks effectively can vary
- Evaluation is crucial, in both Greek and English

Model Evaluation

Evaluating Meltemi

- Evaluating foundation models involves a combination of quantitative and qualitative assessments to ensure they perform effectively across tasks
- Common method and metrics used:
  - Benchmarking on Standard Datasets
  - Quantitative Metrics (Perplexity, Accuracy, BLEU/ROUGE, WER etc.)
  - Human Evaluation and Error Analysis
- To that end we created a standardized evaluation suite for the Greek language, integrated with the lighteval framerwork
- Also evaluated on the OpenLLM Leaderboard tasks for English

The ILSP LLM evaluation suite for Greek

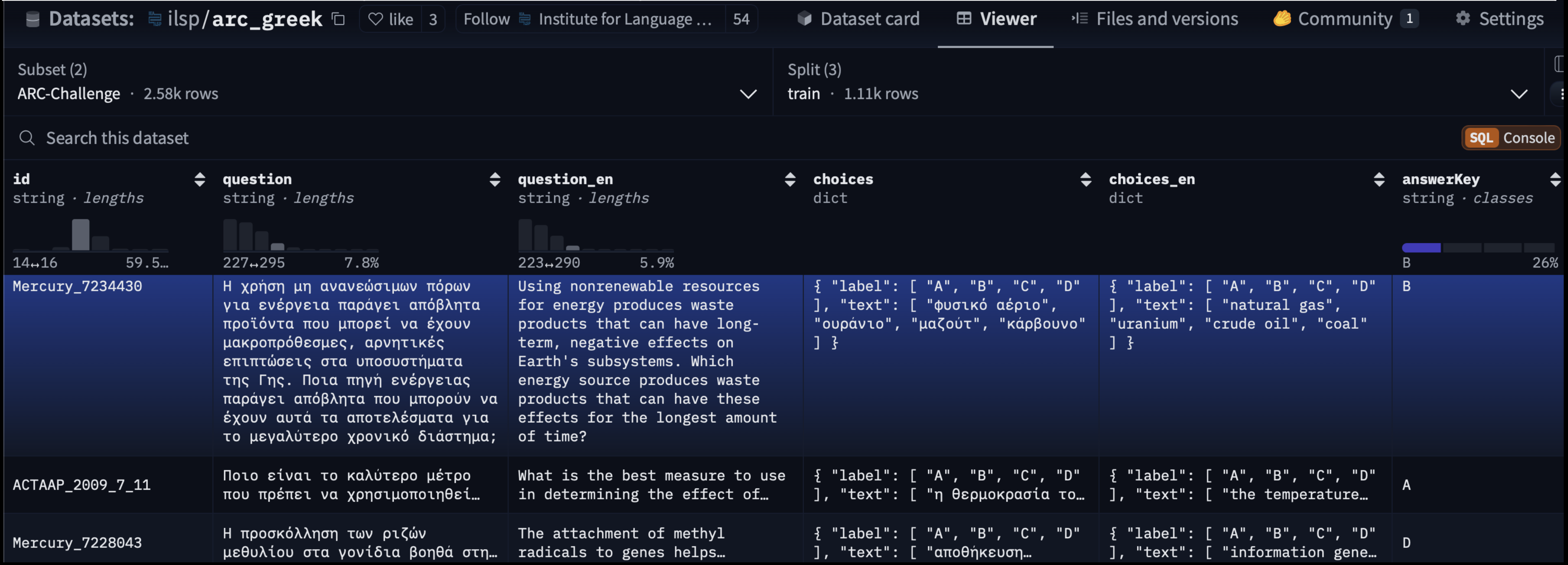
- The evaluation suite comprises of post-edited machine translated versions of publicly available and established English benchmarks for
  - **Language understanding and reasoning**
    - MMLU
    - HellaSwag
    - ARC (2 distinct sets, challenge and easy)
  - **General Question Answering**
    - Truthful QA
    - Winogrande
    - Belebele (8-shot)

The ILSP LLM evaluation suite for Greek (cont.)

- It also contains a novel benchmark with questions extracted from past medical exams (Medical MCQA for Greek)
- All datasets are publicly available through Hugging Face, under https://huggingface.co/ilsp

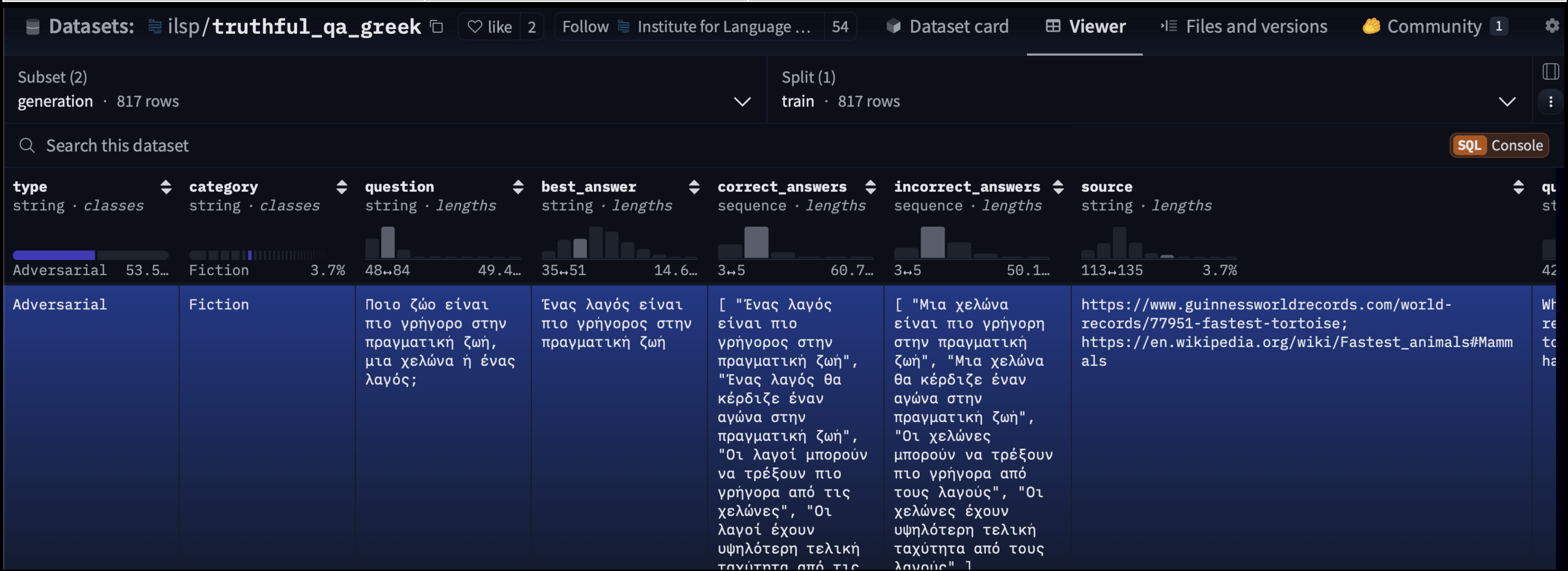
The ILSP LLM evaluation suite for Greek - details

Name	# Examples	Description
ARC Greek	7.78K	MT of ARC (Clark et al., 2018), a dataset of science exam questions (with typically four answer options) partitioned into a Challenge and an Easy Set of 2.6K and 5.2K questions



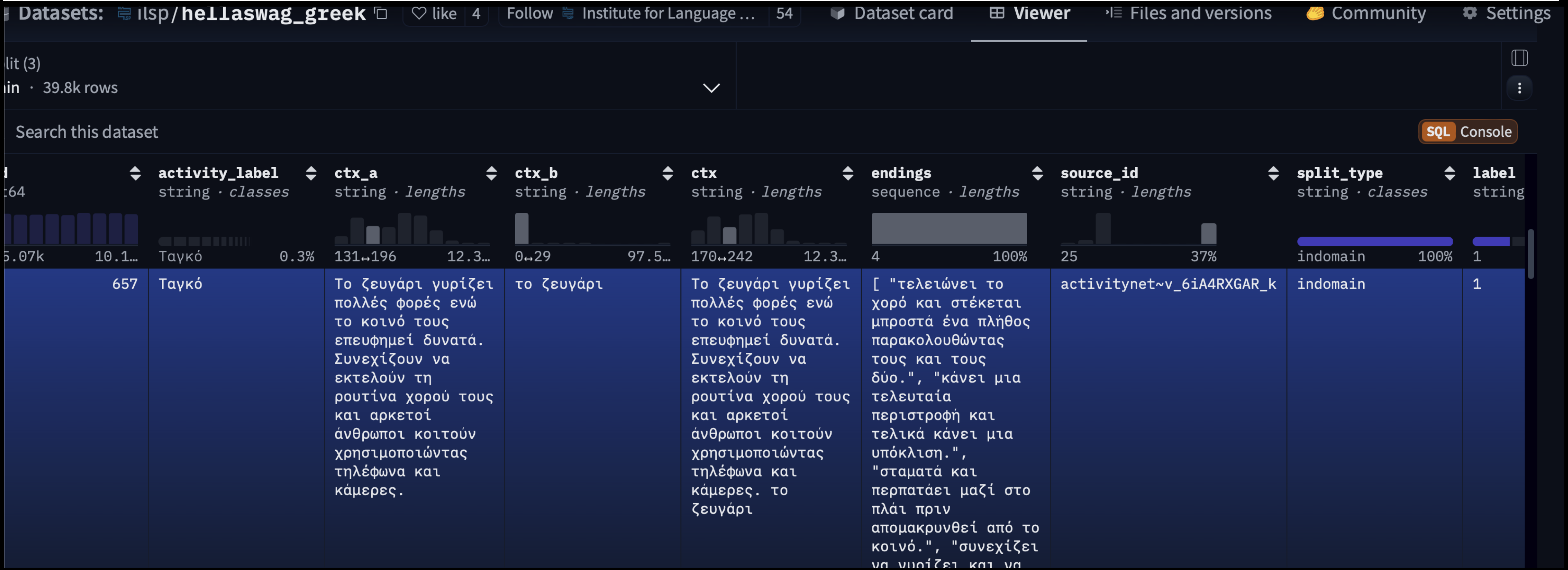
The ILSP LLM evaluation suite for Greek - details

Name	# Examples	Description
Truthful QA Greek	817	Edited MT of Truthful QA (Lin et al., 2022), a dataset of questions that are crafted so that some humans would answer wrongly due to a false belief or misconception.



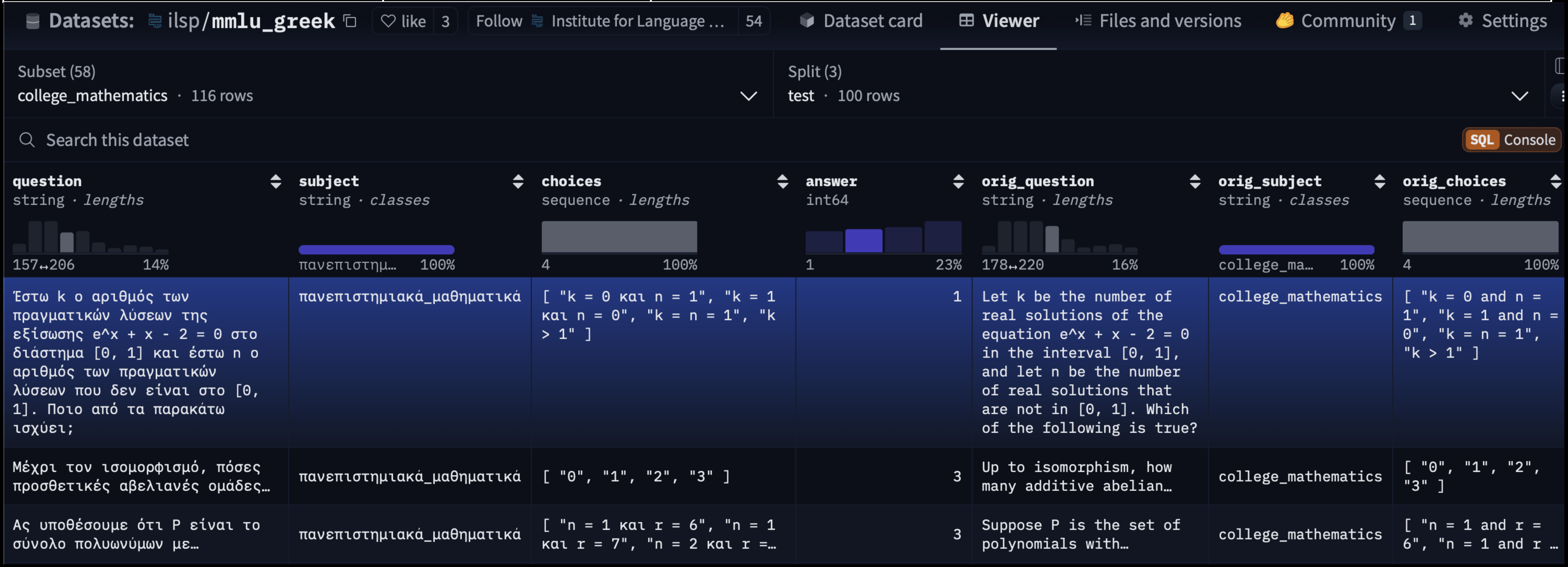
The ILSP LLM evaluation suite for Greek - details

Name	# Examples	Description
HellaSwag Greek	59.8K	MT of the HellaSwag dataset (Zellers et al., 2019) for commonsense NLI



The ILSP LLM evaluation suite for Greek - details

Name	# Examples	Description
MMLU Greek	15.9K	MT of the MMLU dataset (Hendrycks et al., 2021) of multiple-choice questions from 57 tasks including elementary mathematics, history, computer science, law, etc.



Datasets: ilsp/mmlu_greek · like 3 · Follow Institute for Language ... 54 · Dataset card · Viewer · Files and versions · Community 1 · Settings

Subset (58): college_mathematics · 116 rows

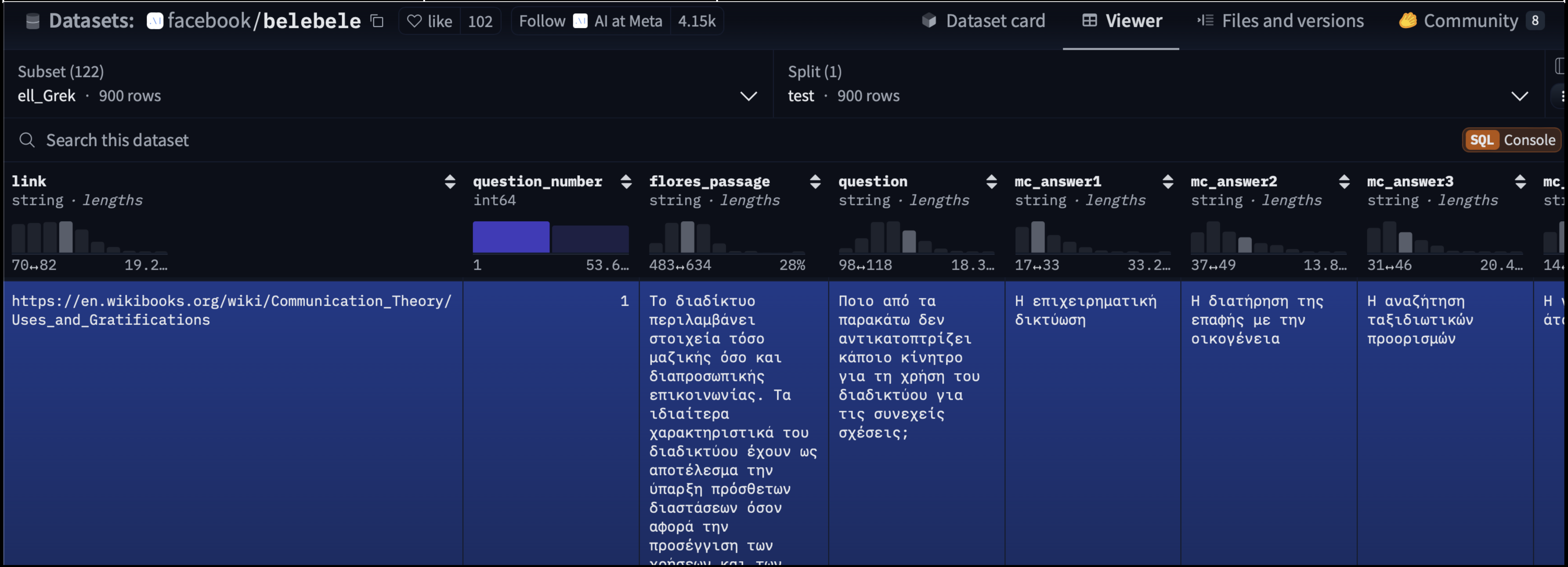
Split (3): test · 100 rows

Search this dataset · SQL Console

question string · lengths 157↔206 14%	subject string · classes πανεπιστημ... 100%	choices sequence · lengths 4 100%	answer int64 1 23%	orig_question string · lengths 178↔220 16%	orig_subject string · classes college_ma... 100%	orig_choices sequence · lengths 4 100%
Έστω k ο αριθμός των πραγματικών λύσεων της εξίσωσης e^x + x - 2 = 0 στο διάστημα [0, 1] και έστω n ο αριθμός των πραγματικών λύσεων που δεν είναι στο [0, 1]. Ποιο από τα παρακάτω ισχύει;	πανεπιστημιακά_μαθηματικά	["k = 0 και n = 1", "k = 1 και n = 0", "k = n = 1", "k > 1"]	1	Let k be the number of real solutions of the equation e^x + x - 2 = 0 in the interval [0, 1], and let n be the number of real solutions that are not in [0, 1]. Which of the following is true?	college_mathematics	["k = 0 and n = 1", "k = 1 and n = 0", "k = n = 1", "k > 1"]
Μέχρι τον ισομορφισμό, πόσες προσθετικές αβελιανές ομάδες…	πανεπιστημιακά_μαθηματικά	["0", "1", "2", "3"]	3	Up to isomorphism, how many additive abelian…	college_mathematics	["0", "1", "2", "3"]
Ας υποθέσουμε ότι P είναι το σύνολο πολυωνύμων με…	πανεπιστημιακά_μαθηματικά	["n = 1 και r = 6", "n = 1 και r = 7", "n = 2 και r =…	3	Suppose P is the set of polynomials with…	college_mathematics	["n = 1 and r = 6", "n = 1 and r …

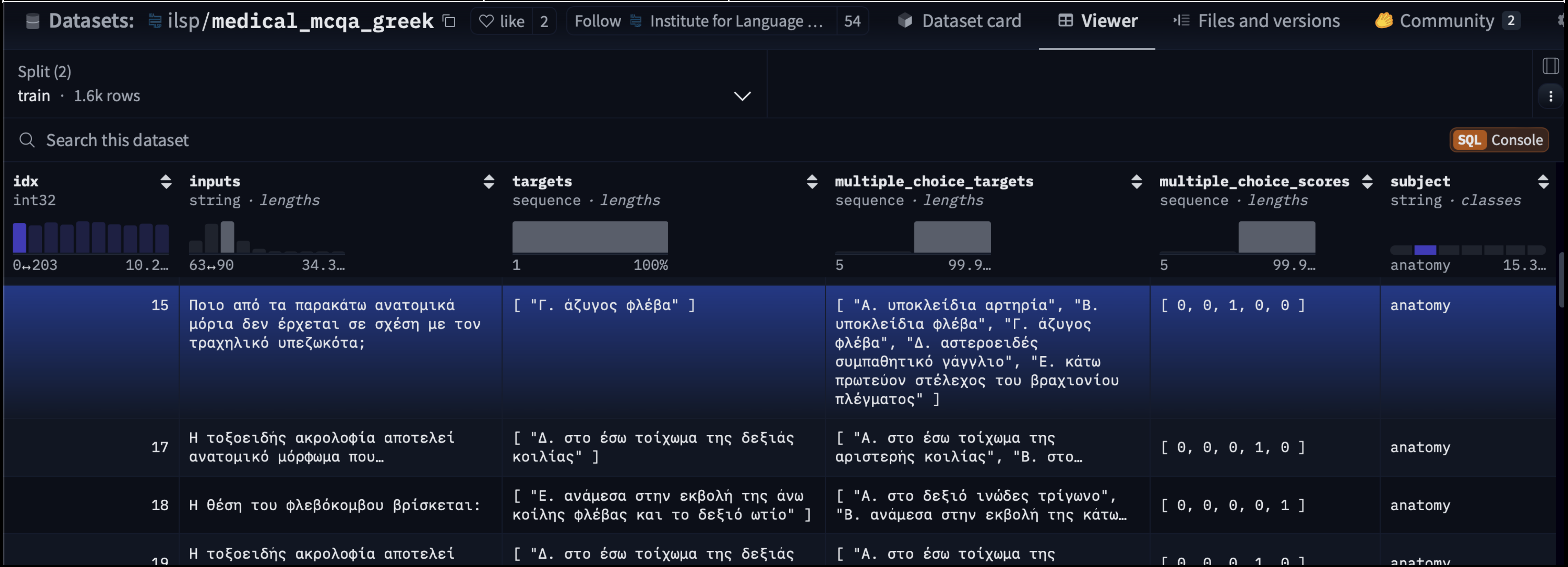
The ILSP LLM evaluation suite for Greek - details

Name	# Examples	Description
Belebele (ell)	900	The Greek part of Belebele (Bandarkar et al., 2023), a multiple-choice machine reading comprehension dataset covering 122 language variants.



The ILSP LLM evaluation suite for Greek - details

Name	# Examples	Description
Greek Medical Multiple Choice QA	2.03K	Multiple choice questions extracted from past medical exams of the Greek National Academic Recognition and Information Center available at https://www.doatap.gr



Model Evaluation for Greek

	Medical MCQA 15-shot	Belebele 5-shot	HellaSwag 10-shot	ARC-C 25-shot	Truthful QA 0-shot	MMLU 5-shot	Avg.
Mistral 7B	27.7%	35.7%	35.2%	27.2%	44.9%	24.8%	32.5%
Meltemi 7B v1	**46.3%**	**68.5%**	**63.3%**	**43.6%**	**44.6%**	**42.4%**	**51.4%**
Meltemi 7B v1.5	**48.1%**	**68.6%**	**65.7%**	**47.1%**	**45.1%**	**42.4%**	**52.8%**

- Our evaluation for Meltemi-7B-v1 and v1.5 is performed in a few-shot setting, consistent with the settings in the Open LLM leaderboard
- Meltemi v1.5 enhances performance across all Greek test sets by a **+20.5%** average improvement.

Model Evaluation for English

	Winogrande	GSM8K	HellaSwag 10-shot	ARC-C 25-shot	Truthful QA 0-shot	MMLU 5-shot	Avg.
Mistral 7B	**78.37%**	**34.5%**	**83.31%**	**59.98%**	**42.15%**	**64.16%**	**60.4%**
Meltemi 7B v1.5	73.1%	22.1%	79.6%	54.2%	40.6%	56.8%	54.4%

- The model performs worse than Mistral-7B for the English tasks, trailing by **-6%**
- The difference in style between Greek and English data the model was initially trained on, has impacted the performance in English tasks

Comparing our results with similar efforts

- **LeoLM-7B**: A German foundation model created by applying continual pretraining on Llama2-7B
  - German benchmarks: average improvement of **+4.8%**
  - English benchmarks: achieves lower average scores by **-2.6%**
- **Swallow-MS-7b-v0.1**: A Japanese foundation model created by applying continual pretraining on Mistral-7B
  - Japanese benchmarks: average improvement of +8%
  - English benchmarks: achieves lower average scores by **-5.3%**

Creating a Chat Model

Creating a chat model

- The base model is trained for text completion
  - Given the past words predict the next
  - Not able to answer user questions by itself
- For the chat we need an instruction following model
  - Trained using pairs of [query, response] using Supervised Fine-Tuning (SFT)
- Alignment for human preference tuning with triplets of (query, good response, bad response)

Instruction data

- ~100K instruction pairs used for training
- Obtained from public corpora and translated to Greek using state-of-the-art MT models
  - Open-Platypus (subsets with permissive license)
  - Evol-Instruct
  - Capybara
- Example (from Open-Platypus)
  - Instruction: Βρες με πόσους τρόπους μπορούν να αναδιαταχτούν τα γράμματα της λέξης TARTAR
  - Response: Για να αναδιατάξω τα γράμματα μιας λέξης, μπορώ να χρησιμοποιήσω τον τύπο για μεταθέσεις n αντικειμένων, όπου μερικά είναι πανομοιότυπα. Σε αυτή την περίπτωση, n = 6, ...

Chat model training

- Fine-tuning has been done with the Odds Ratio Preference Optimization (ORPO) algorithm using 97k preference data:
  - 89,730 Greek preference data which are mostly translated versions of high-quality datasets on Hugging Face
  - 7,342 English preference data
- Alignment procedure is based on the TRL - Transformer Reinforcement Learning library and partially on the Hugging Face finetuning recipes

Safety and Handling Harmful content

- Automatic translation of English safety instructions from pre-existing datasets
- Manual creation of ~100 Greek custom safety instructions to limit harmful responses
- Manual inspection of model's responses for various unsafe cases (e.g., creation of illegal substances/bombs, self-harm guidance, explicit content, inappropriate jokes, etc.)

Evaluating the chat model for Greek

	Medical MCQA 15-shot	Belebele 5-shot	HellaSwag 10-shot	ARC-C 25-shot	Truthful QA 0-shot	MMLU 5-shot	Avg.
Mistral 7B	27.7%	35.7%	35.2%	27.2%	44.9%	24.8%	32.5%
Meltemi 7B v1.5	48.1%	68.6%	**65.7%**	**47.1%**	45.1%	42.4%	52.8%
Meltemi 7B Chat v1.5	**46.5%**	**76.8%**	64.7%	46.5%	**54.2%**	**45.4%**	**55,6%**

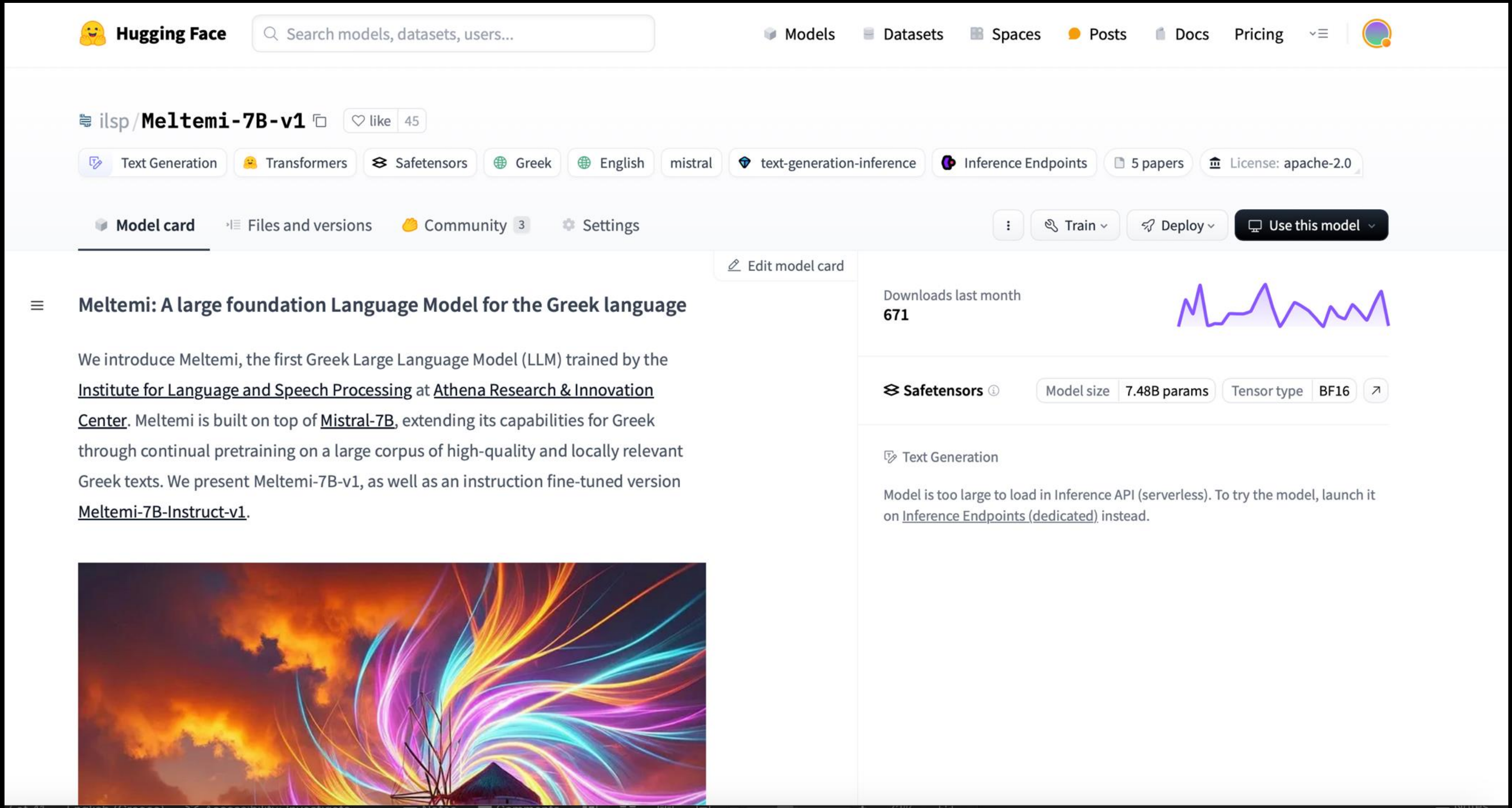
- Meltemi Chat enhances performance across all Greek test sets by a **+2.8%** average improvement over the foundation model.

How to use Meltemi

How to use Meltemi

1. Download the model directly from Hugging Face
2. As an API endpoint

Download the model directly from Hugging Face



Download the model directly

- We have uploaded on Hugging Face various versions of the model:
  - Meltemi-7B-v1 & v1.5: The foundation model (2 versions)
  - Meltemi-7B-Instruct-v1 & v1.5: The chat model base on Meltemi-7B-v1
  - Multiple quantized versions of the instruct model
    - Bits and Bytes 4-bit: fast version
    - AWQ 4-bit: slower but may have slightly better performance
    - GGUF versions: ready to deploy using llama.cpp / ollama
- On each model card information of how one can use the model with various Python libraries, such as transformers, llama_cpp and awq

As an API endpoint

- You can access Meltemi through an API endpoint
- We currently provide access to the model:
  - Using the OpenAI client API
  - Using the API

Accessing Meltemi using the OpenAI client API

- We provide access to the model through a proxy so that you can call it using the OpenAI client API
- For model access you will need an API key:

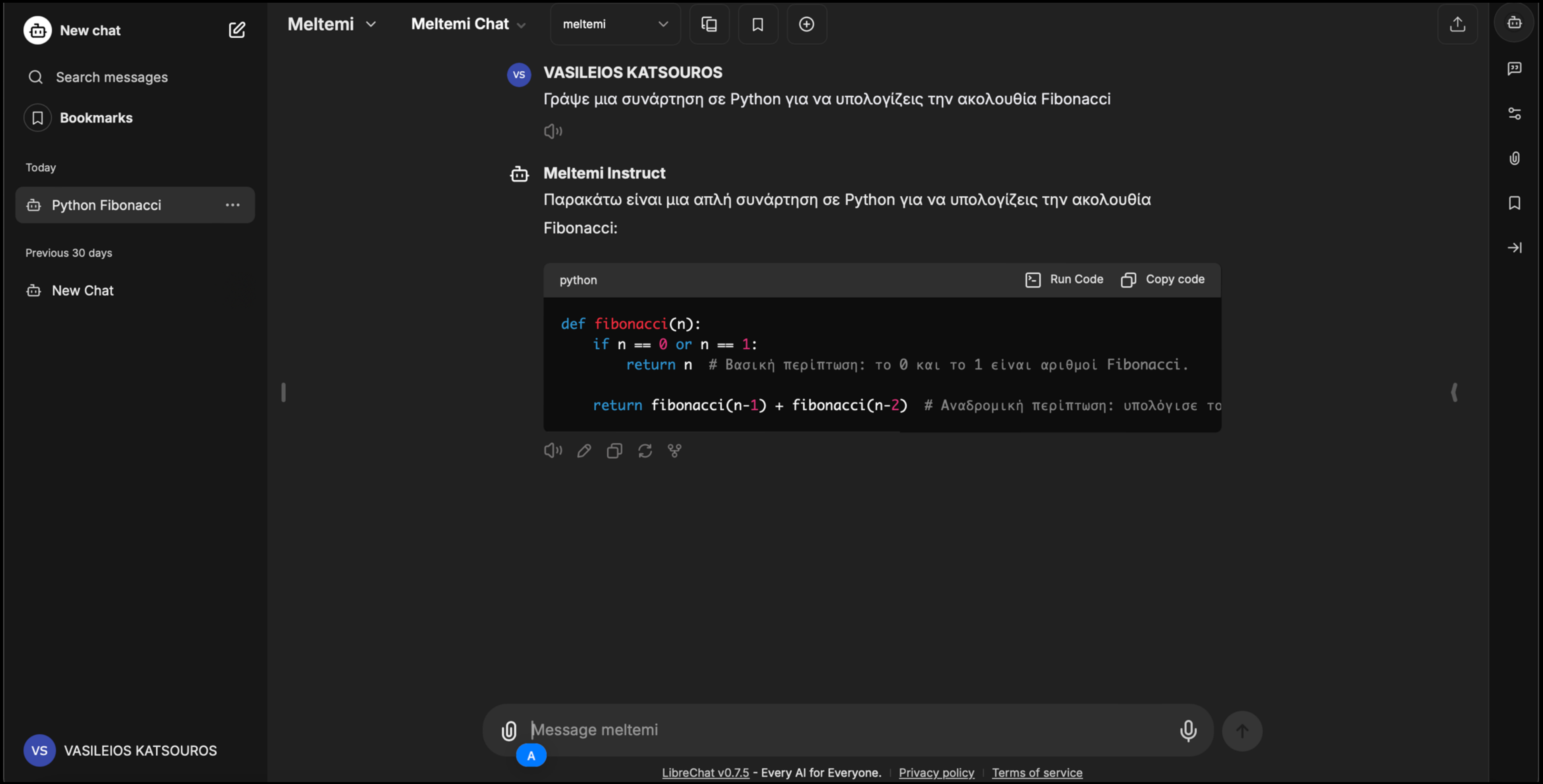
```python
from openai import OpenAI
MELTEMI_API_KEY = "sk--I0Ld3h6yeH1YOGimVmJ6g"
MELTEMI_BASE_URL = "http://ec2-3-19-37-251.us-east-2.compute.amazonaws.com:4000/"

MELTEMI_CLIENT = OpenAI(api_key=MELTEMI_API_KEY, base_url=MELTEMI_BASE_URL)
query = "Η Αλίκη έχει 5 αδερφές και 5 αδερφούς. Πόσες αδερφές έχει ένας αδερφός της Αλίκης;"
system_prompt = "Είσαι."
response = MELTEMI_CLIENT.chat.completions.create(model = "meltemi",
messages = [{"role": "system", "content": system_prompt},
            {"role": "user", "content": query},
           ],)
print(response.choices[0].message.content)
```

Accessing Meltemi using the API

- You can send requests to our API
  - curl --location "http://ec2-3-19-37-251.us-east-2.compute.amazonaws.com:4000/chat/completions" --header "Authorization: Bearer sk--I0Ld3h6yeH1YOGimVmJ6g" --header "Content-Type: application/json" --data "{\"model\": \"meltemi\", \"messages\": [{\"role\": \"system\", \"content\": \"Είσαι το....\"}, {\"role\": \"user\", \"content\": \"Η Αλίκη έχει 5 αδερφές και 5 αδερφούς. Πόσες αδερφές έχει ένας αδερφός της Αλίκης;\"}]}"

Accessing Meltemi using Open WebUI

- setup Open WebUI to chat with the model through a web interface http://meltemi.ilsp.gr



Outcomes

Outcomes

- Released all models with Apache 2.0 license on Hugging Face
  - Two model variants:
    - Foundation Model: Meltemi-7B-v1 & v1.5
    - Chat Model: Meltemi-7B-Instruct-v1 &v1.5
  - Quantized versions to run locally
- Created evaluation suite with 6 test sets for Greek, also shared with the research community on Hugging Face
- Access Meltemi API
- Chat with Meltemi http://meltemi.ilsp.gr

Next steps

Next Steps

- Gather more data resources
- Expanding our models' capabilities for:
  - Translation (EN-EL & EL-EN)
  - Instruction-following and chats
  - Creating synthetic Greek tasks from existing data
  - RAG (Retrieval-Augmented Generation) applications
  - Function calling agents
- LLama3.1

Acknowledgements

grnet

OCRE
Open Clouds for Research Environments

Meltemi: The first open Large Language Model for Greek

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos,
Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou,
Athanasios Katsamanis, Stelios Piperidis, Vassilis Katsouros

Institute for Speech and Language Processing, Athena Research Center
Artemidos 6 & Epidavrou, Athens, Greece
vsk@athenarc.gr

Thank you!