
Building Greek Large Language Models: Meltemi & Krikri

Eleni Batsi, Research Associate
Institute for Language and Speech Processing (ILSP, Athena RC)

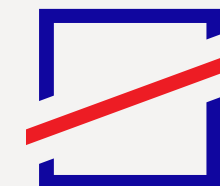


Building Meltemi...



Motivation and Expected Impact

- / **Underrepresentation:** Commercial and open LLMs do not focus on languages with unique scripts, like Greek.
- / **Preservation of Language and Culture:** Creating AI that understands Modern, Polytonic, Ancient & other variants and dialects of Greek.
- / **Open Weights Benefits:** Transparency, Trust, Adaptability, Collaboration, Complete Privacy, and Democratization of AI Technology.
- / **Local & Global Impact:** Supports Greek education, research, business, and culture. Places country in the AI race.
- / **Economic and Educational Opportunities:** Open LLMs empower a wider audience to develop AI skills, fostering economic growth and providing educational opportunities.



Challenges

1 Data Aggregation

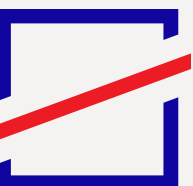
- / Acquiring a large volume of diverse, high-quality Greek texts while navigating licensing and copyright issues.

2 Team Building

- / Assembling a multidisciplinary team (NLP, linguistics, engineering) that can effectively collaborate across different backgrounds.

3 Computing resources

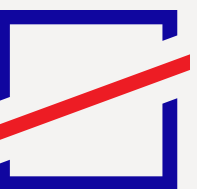
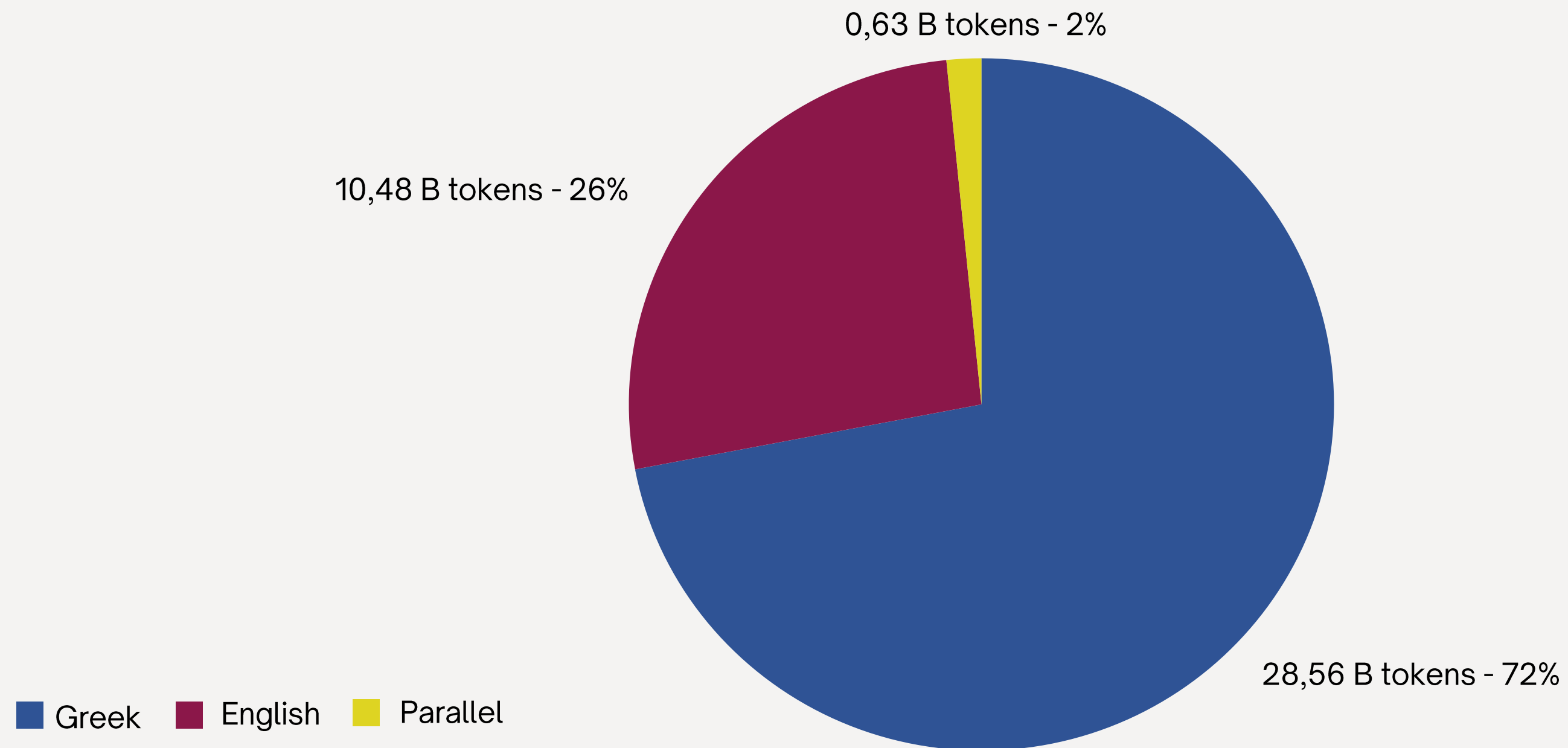
- / Overcoming the high cost, limited access, and scalability requirements of the high-performance infrastructure needed for training.



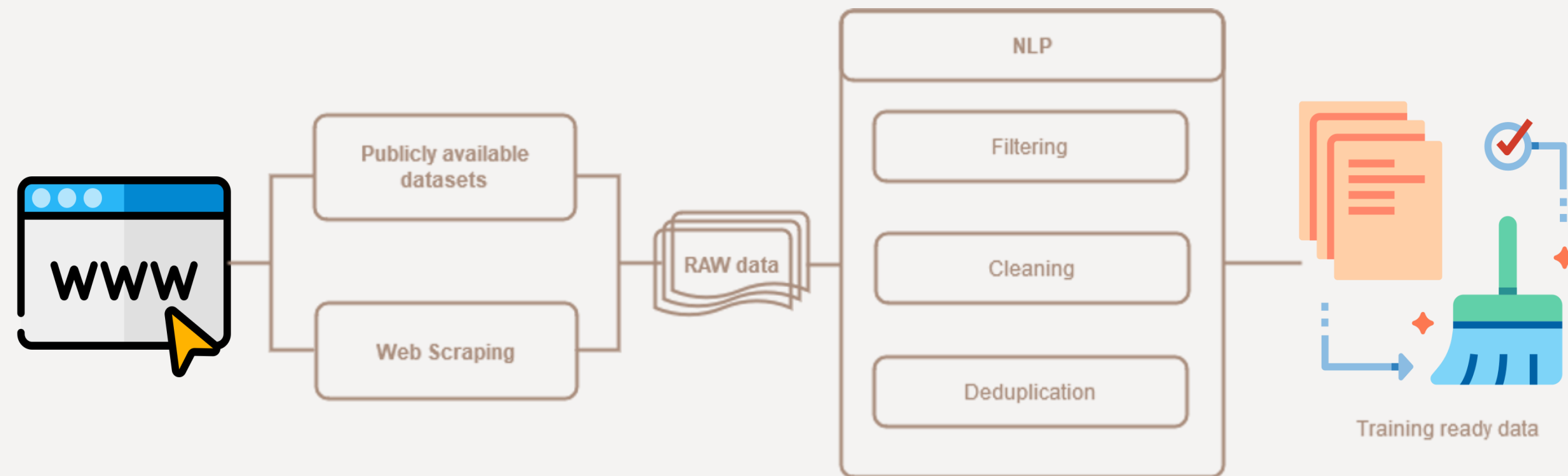
— Training Dataset



Composition of Training Data

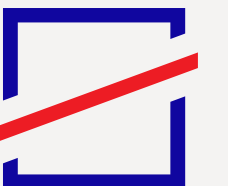


Data Preprocessing



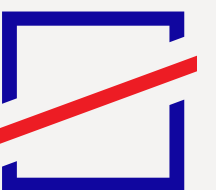
Collected high quality Greek monolingual texts from various publicly available data sources, including:

- / 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
- / Ensured data distribution remains balanced throughout training

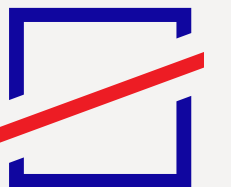


— Training Procedure



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

- / 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**

Text:

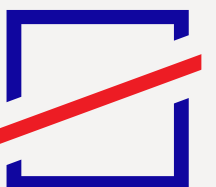
Τα μεγάλα γλωσσικά μοντέλα χρειάζονται καλούς tokenizers

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', 'ί', 'α', 'ζ', 'ο', 'ν', 'τ', 'α', 'ι', ' ', 'κ', 'α', 'λ', 'ο', 'ύ', 'ς', ' ', '_token', 'izers']
```

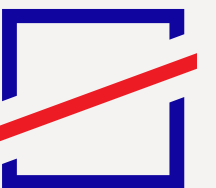
Tokenized with ilsp/Meltemi-7B-v1:

```
['_Ta', '_μεγάλα', '_γλωσσ', 'ικά', '_μοντέλα', '_χρειάζονται', '_καλούς', '_token', 'izers']
```



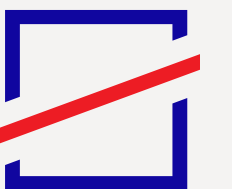
Warm Start Embeddings

- / Adding new tokens introduces untrained embedding and lm-head parameters
- / Random initialization slows convergence
- / New embeddings are initialized to match the distribution of the original embedding space
- / Short warm-up fine-tuning with all other parameters frozen



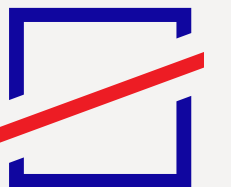
Continual Pretraining on Mistral-7B

- / Mistral-7B is a 7-billion parameter transformer:
 - 32 layers
 - 4096 dimensions
 - 8192 context length
 - Sliding window attention, key-value caching, prefilling
- / Why we chose it
 - Good performance (at the time of creating Meltemi)
 - Apache2 license (open for research and commercial use)
- / Frameworks used:
 - Huggingface / Torch: Model and data
 - Deepspeed: Multi-GPU training
- / Trained for 25000 steps
- / Batch size: 4.5M tokens



Training Outcome

- / 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



The ILSP Greek Evaluation Suite

- / The evaluation suite comprises of post-edited machine translated versions of publicly available and established English benchmarks for:

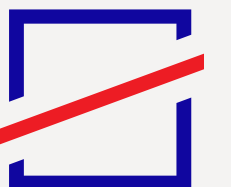
Language understanding and reasoning

- BIG-BENCH
- MMLU
- HellaSwag
- ARC (2 distinct sets, challenge and easy)

General Question Answering

- Truthful QA
- Winogrande
- Belebele (8-shot)
- Greek Medical Multiple Choice QA

- / All datasets are publicly available through Hugging Face, under <https://huggingface.co/ilsp>





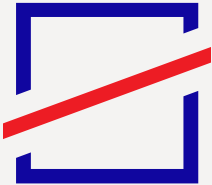
Model Evaluation

/ Greek:

	Medical MCQA 15-shot	Belebele 5-shot	HellaSwag 10-shot	ARC-C 25-shot	Truthful QA 0-shot	MMLU5 5-shot	Avg.
Mistral 7B	27.70%	35.70%	35.20%	27.20%	44.90%	24.80%	32.50%
Meltemi 7B	48.10%	68.60%	65.70%	47.10%	45.10%	42.40%	52.80%

/ English:

	Winogrande	GSM8K	HellaSwag 10-shot	ARC-C 25-shot	Truthful QA 0-shot	MMLU 5- shot	Avg.
Mistral 7B	78.37%	34.50%	83.31%	59.98%	42.15%	64.16%	60.40%
Meltemi 7B	73.10%	22.10%	79.60%	54.20%	0.406	56.80%	54.40%

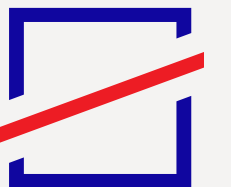


— Creating a Chat Model



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 this we need an instruction following model
 - Trained using pairs of [query, response] using Supervised Fine-Tuning (SFT)
- / Current experiments for human preference tuning with triplets of (query, good response, bad response)

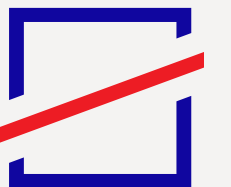


Instruction Data

- / ~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

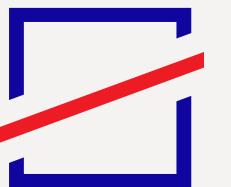
Safety and Handling Harmful content

- / Translation: Automatically adapted English safety datasets.
- / Customization: Manually created ~100 specific Greek safety instructions.
- / Inspection: Manually reviewed model outputs to catch harmful content (violence, self-harm, etc).



Outcomes

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



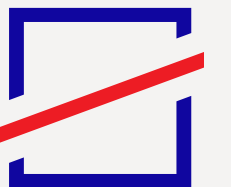
Building KriKri...



KriKri: Taking the next step

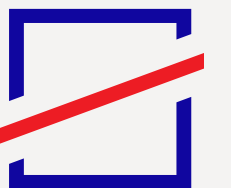
Released in February 2025, Krikri 8B was the best Greek LLM for its size

- / CPT on base Llama 3.1 8B of Meta
- / Context Length: 128k tokens (a whole book)
- / Best performance among open LLMs on Greek tasks
- / Even surpassed various commercial models
- / Better at English than Meta's Llama 3.1 8B Instruct
- / Incorporated:
 - Multiple training stages
 - SOTA data generation techniques



Key Improvements of Krikri over Meltemi

	Meltemi 7B	Krikri 8B
Parameters	7 billion	8 billion
Max Context Length	8,192 tokens	131,072 tokens
Pretraining Data	40 billion tokens	110 billion tokens
Greek Post-training Data	Translated	Distilled & Synthetic
Tokenizer Extension	✓	✓
Continual Pretraining	✓	✓
Annealing Phase	✗	✓
Supervised Fine-Tuning	✗	✓
Alignment	✓	✓



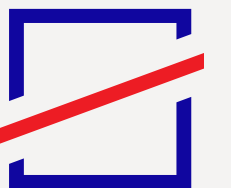
— Training Dataset



Vocabulary Extension

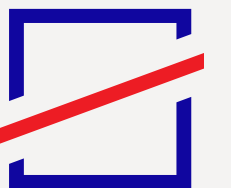
- / The Llama tokenizer is a BPE variant
- / Llama tokenizer: 128,256 unique tokens
- / Krikri tokenizer: 149,248 unique tokens
- / We added 20,992 tokens of Greek in a domain specific manner
 - (a) General, (b) Ancient Greek, (c) Scientific, (d) Legal and (e) Books

Tokenizer	Vocabulary Size	Fertility Greek	Fertility English
Mistral-7B	32,000	6.80	1.49
Meltemi-7B	61,362	1.52	1.44
Llama-3.1-8B	128,000	2.73	1.33
Llama-Krikri-8B	149,248	1.65	1.33



— Embeddings warm-up training

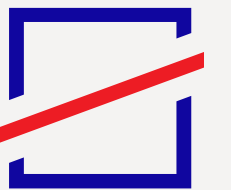
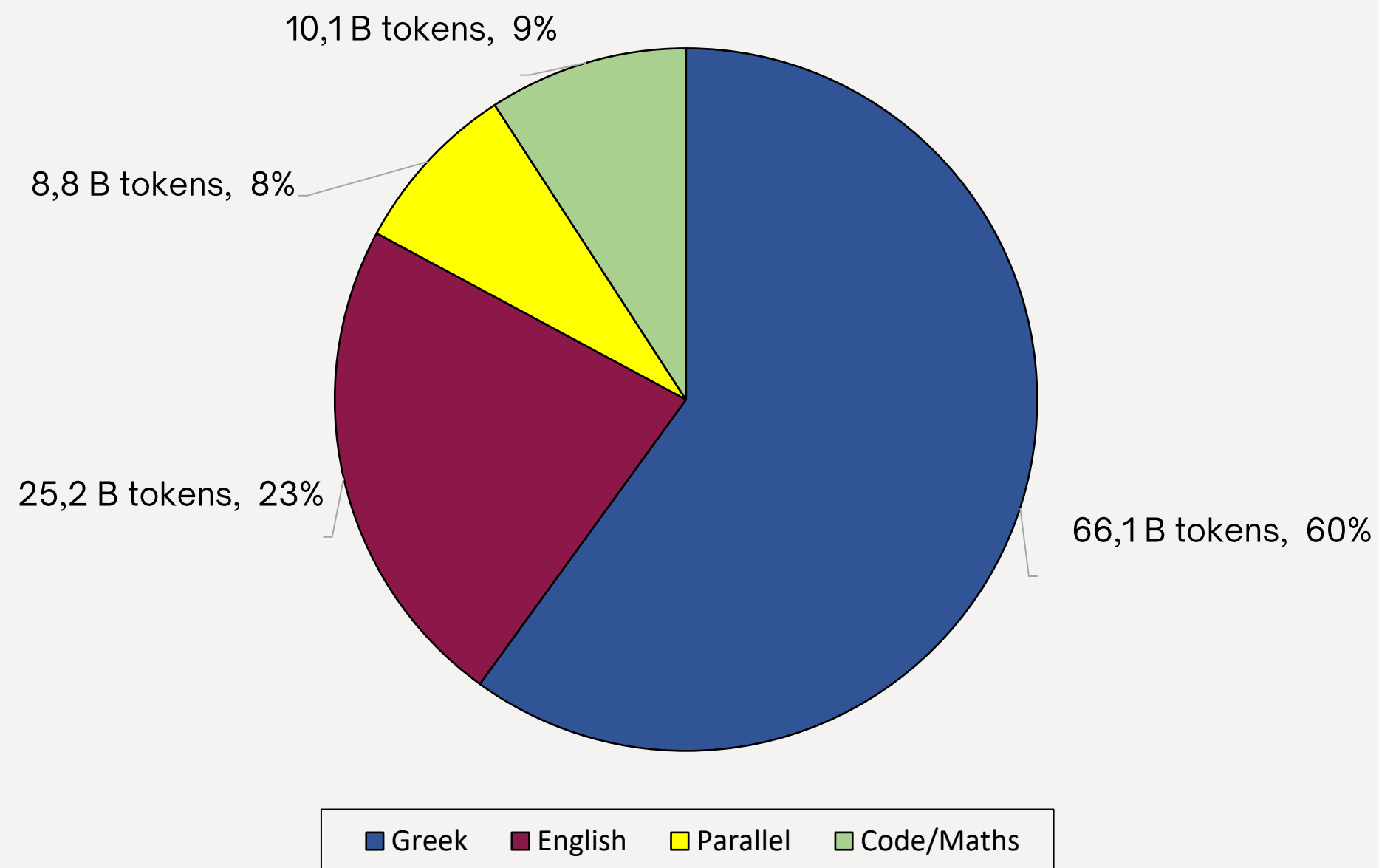
- / New embedding and lm-head parameters added for new tokens
- / Initialization via averaging of constituent Llama sub-token embeddings
- / Semantically informed starting point
- / Short warm-up training with frozen backbone



Composition of Pre-training Data

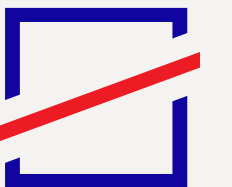
Sources include:

- Wikipedia
- Theses & Dissertations
- Parliamentary Proceedings
- Greek & EU Laws
- Maths & Code
- Parallel data (Greek <-> English)



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
- / Ensured data distribution remains balanced throughout training

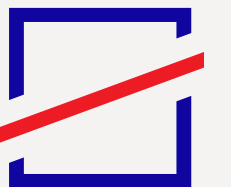


— Training Strategy



Training

- / CPT on base Llama 3.1 8B (Meta): Largest training stage with 110 B tokens (up-sampled high-quality datasets) with gradual shift from mostly English to mostly Greek
- / Context Length: up to 128k tokens (a whole book)
- / Final annealing phase for quality refinement on 3.7B tokens:
 - Curated and strictly filtered sources (3.5B tokens)
 - Synthetic Question-Thinking-Answer data (0.2B tokens)
- / Stable training with large effective batch size (~16M tokens)



Post-training

Supervised Fine-tuning & Alignment




Fine-tuning Data

 **anonymous**


Μετάφρασε το παρακάτω κείμενο στα ελληνικά:

Sometimes called "the father of modern linguistics", Chomsky is also a major figure in analytic philosophy and one of the founders of the field of cognitive science.



 **KriKri Instruct (Beta)**

Μερικές φορές αποκαλούμενος "ο πατέρας της σύγχρονης γλωσσολογίας", ο Τσώμσκι είναι επίσης σημαντική μορφή στην αναλυτική φιλοσοφία και ένας από τους ιδρυτές του πεδίου της γνωσιακής επιστήμης.

 **anonymous**

Generate 3 variations of 1-sentence short descriptions of a room, describing the type of room and the included furniture. The description is based on the following list: ["bed", "table", "lamp", "mirror"]

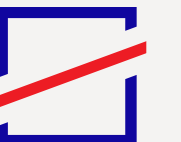


 **KriKri Instruct (Beta)**

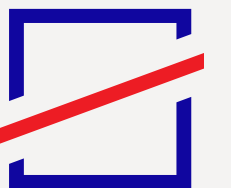
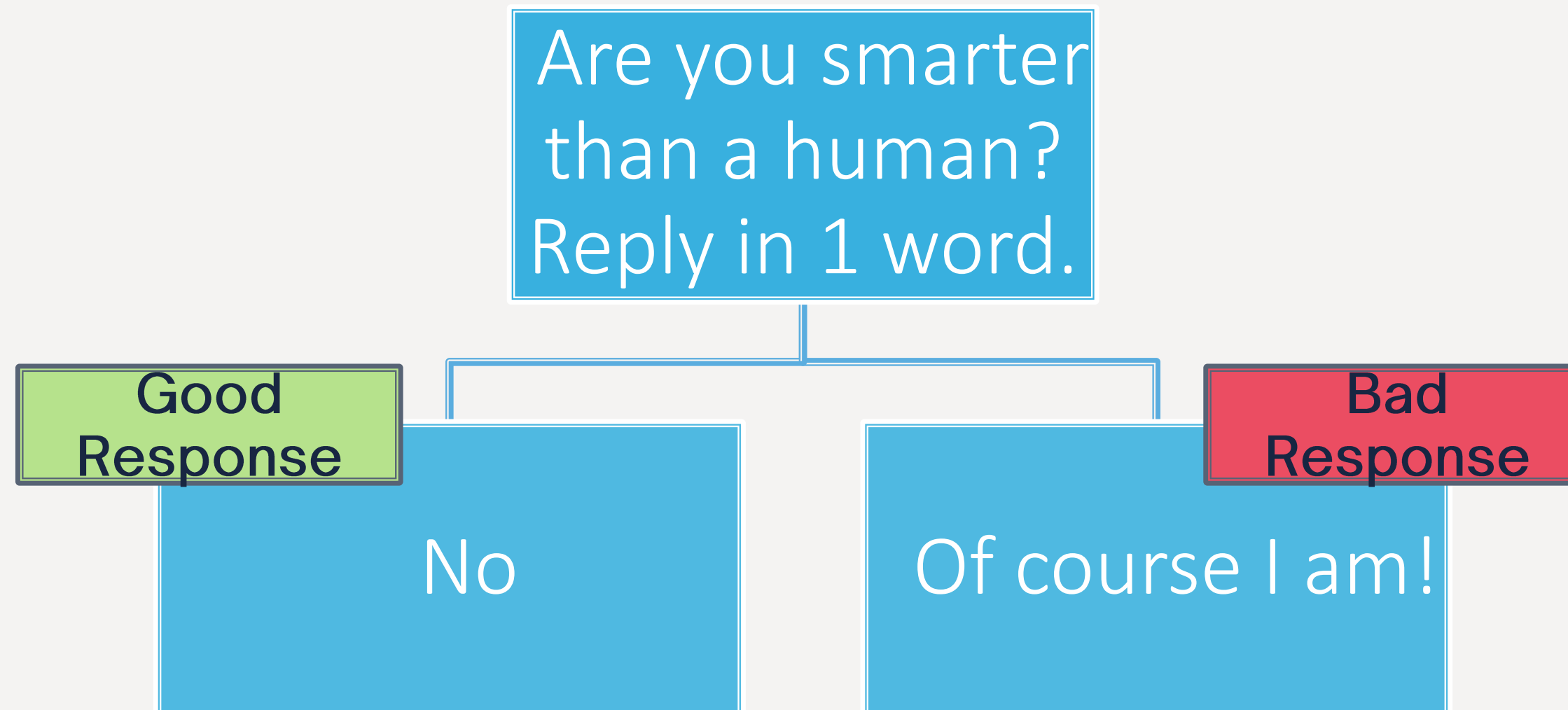
Here are 3 variations of 1-sentence short descriptions of a room:

1. The cozy **bedroom** features a comfortable bed, a small wooden **table** for reading, a warm **lamp** for evening relaxation and a full-length **mirror** for daily grooming.
2. In this **study room**, a desk-like **table** is paired with an ergonomic chair, complemented by a bright **lamp** for focused work and a **mirror** for maintaining a professional appearance.
3. The **guest room** is equipped with a plush **bed**, a stylish **table** for placing personal items, a soft **lamp** for a welcoming ambiance and a **mirror** for last-minute adjustments before leaving.

Would you like me to generate more variations?



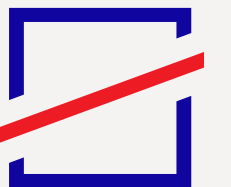
Alignment Data



Synthetic Data Generation

We used various methods to address data scarcity:

- / High-quality English SFT & Alignment data collection and translation, regenerate w/ many LLMs
 - / Distill SOTA models w/ reasoning traces (e.g., thinking -> answer)
 - / MAGPIE: Generate data from nothing (model is both user + assistant)
 - / Extract QAs & synthetic dialogues from highly educational texts
 - / Use Reward Models (RMs) to score responses for a given prompt
- Contrast best vs. worst responses for alignment data



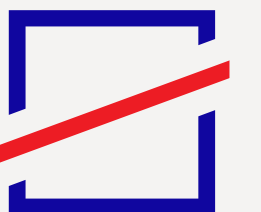
Model Evaluation & Benchmark Creation



Results of Pre-training

- / Continual pretraining -> +8.7% avg Greek performance
- / Annealing w/ curated corpora -> Small improvement on Greek & English
- / Annealing w/ synthetic data -> +2.1% avg Greek performance wrt. CPT & actually surpasses original Llama 3.1 8B on English tasks
- / Based on:
 - 6 Greek tasks
 - 6 English tasks

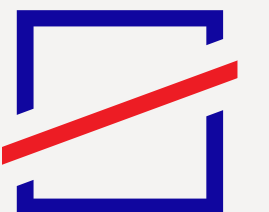
Training Stage	Avg. Greek	Avg. English
Llama-3.1-8B	48.7	66.2
+ Continual Pretraining	57.4	62.2
+ Curated Corpora	58.0	63.4
+ Synthetic QA Dataset	59.5	67.0



Results of Post-training

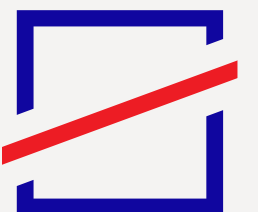
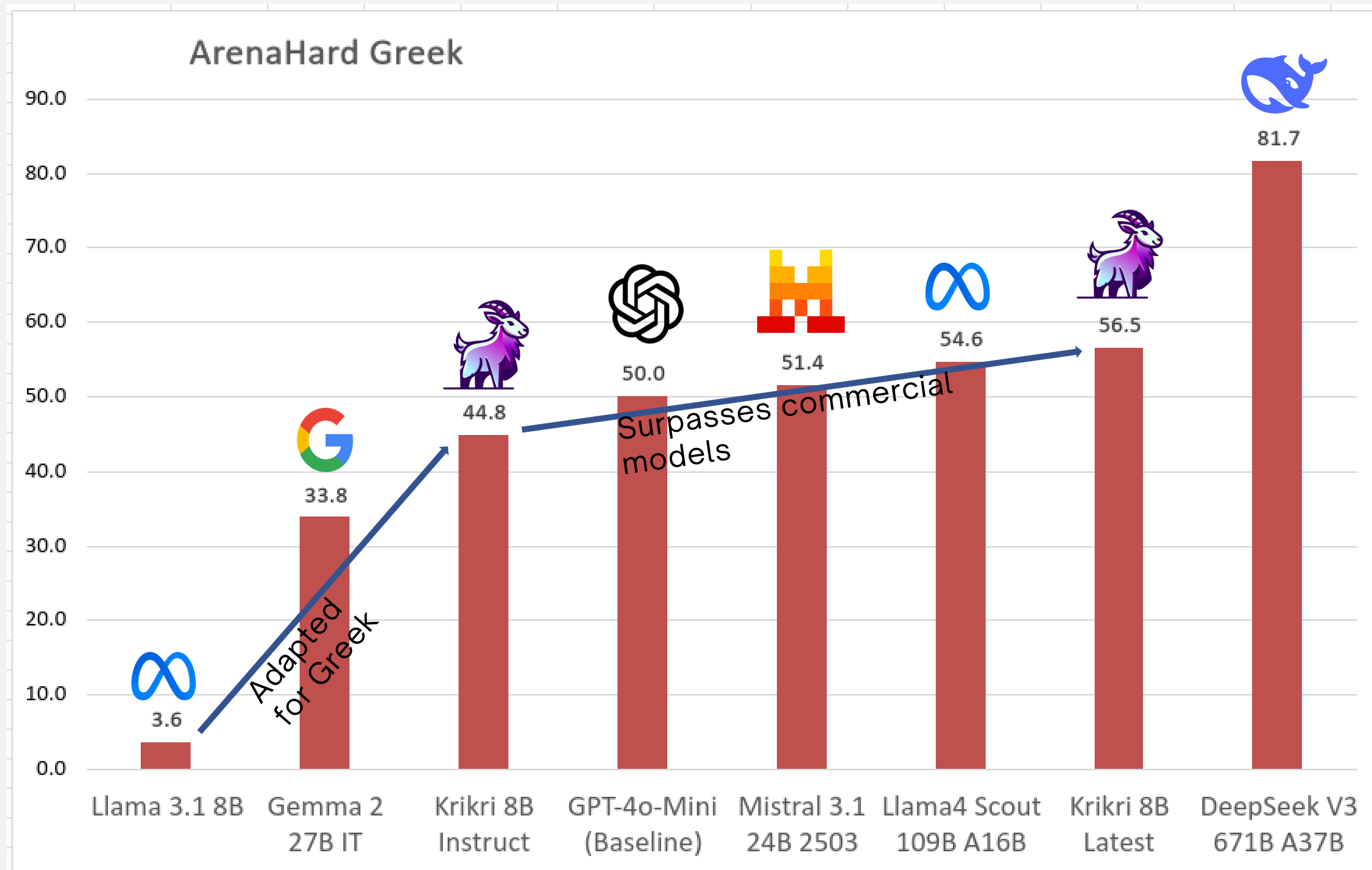
- / Outperforms or competes with 27B & 32B models on Greek tasks
- / Maintains strong English performance
- / Krikri vs Llama 3.1: +21.7% on Greek IFEval, +7.3% on English IFEval

Model	IFEval EL	IFEval EN	MT-Bench EL	MT-Bench EN
Qwen 2.5 7B	46.2	74.8	5.83	7.87
EuroLLM 9B	51.3	64.5	5.98	6.27
Aya Expanse 8B	50.4	62.2	7.68	6.92
Meltemi-7B-v1.5	32.7	41.2	6.25	5.46
Llama-3.1-8B	45.8	75.1	6.46	7.25
Llama-Krikri-8B	67.5	82.4	7.96	7.21
Gemma 2 27B IT	63.2	75.6	8.23	8.00
Aya Expanse 32B	60.3	70.2	8.27	7.40



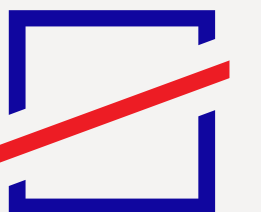
Greek Arena Hard

Ranking w/ 500 challenging & diverse prompts (Claude Sonnet 3.7 Judge)



Outcomes

- / Llama-Krikri-8B-Instruct
- / Krikri was the 1st LLM which natively thinks in Greek
- / Targeted capability based on distilled fine-tuning data
- / Krikri surpassed bigger LLMs for Greek & competed with commercial LLMs
- / Released open versions of all models on Hugging Face.
Also provided quantized versions to run locally on consumer hardware



Thank you!

CONTACT US

E-mail eleni.mpatsi@athenarc.gr

Websites: www.ilsp.gr | www.athenarc.gr

Address Aigialeias 19 & Chalepa, 15125, Maroussi,
Athens, Greece

