

Developing Open Large Language Models for Greek: The Case of Llama-Krikri

Leon Voukoutis & Dimitris Roussis,

Sokratis Sofianopoulos, George Paraskevopoulos, Prokopis Prokopidis, Vassilis Katsouros

Institute for Language and Speech Processing, "Athena" Research Center

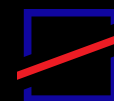


Athens Natural Language Processing Summer School 2025

September 4-10, 2025 - NCSR "Demokritos"

What's in the presentation

- The Journey Towards Greek LLMs
- Multi-stage Training of Krikri
- Data Engineering
- Evaluation & Benchmark Creation
- Next Steps & Frontiers



The Journey Towards Greek LLMs

Athens Natural Language Processing Summer School 2025
September 4-10, 2025 - NCSR "Demokritos"

Motivation

- **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.



Meltemi: Our 1st effort

Released in September 2024, **Meltemi 7B** was the first open Greek LLM

- Based on **Mistral 7B** of Mistral AI
- Context Length: 8k tokens (an academic article)
- Challenges
 - Assembling a team with diverse skills
 - Limited data availability for Greek
 - Tokenizer adaptation
 - Data mix to avoid catastrophic forgetting



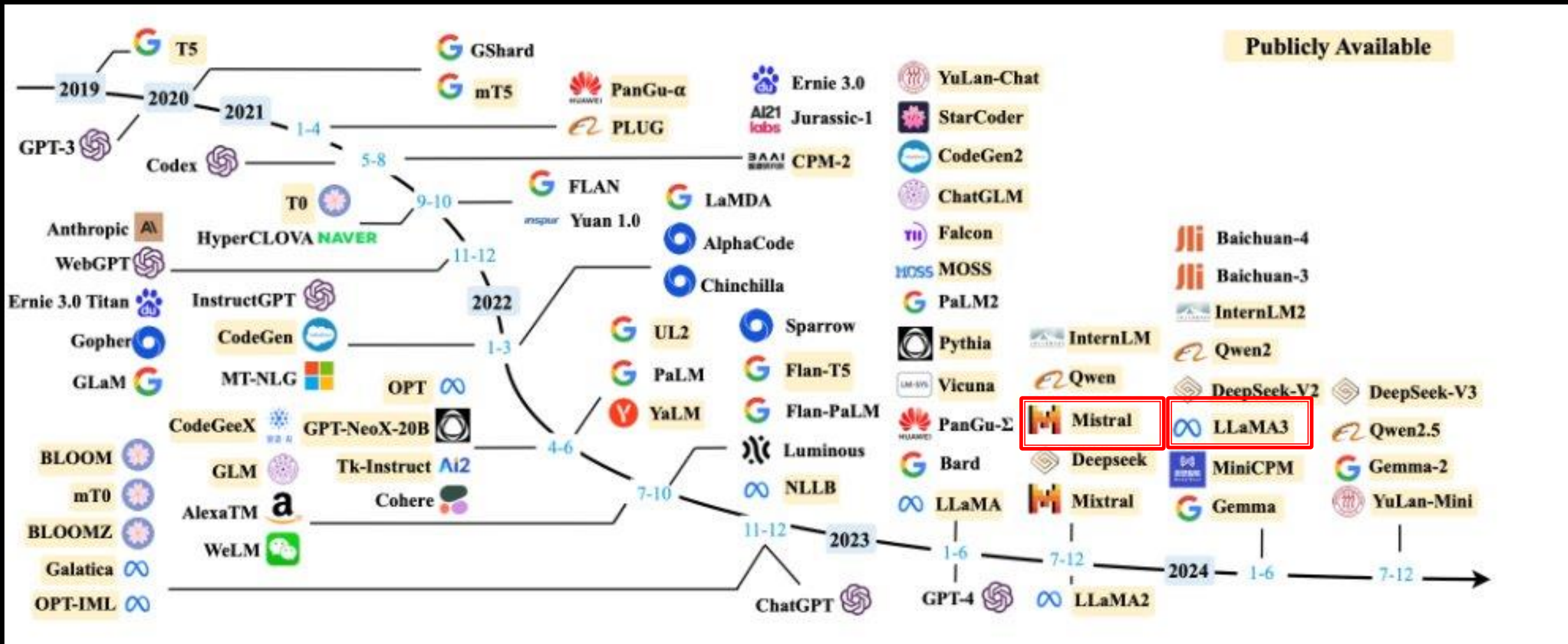
Krikri: Taking the Next Step

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



Timeline



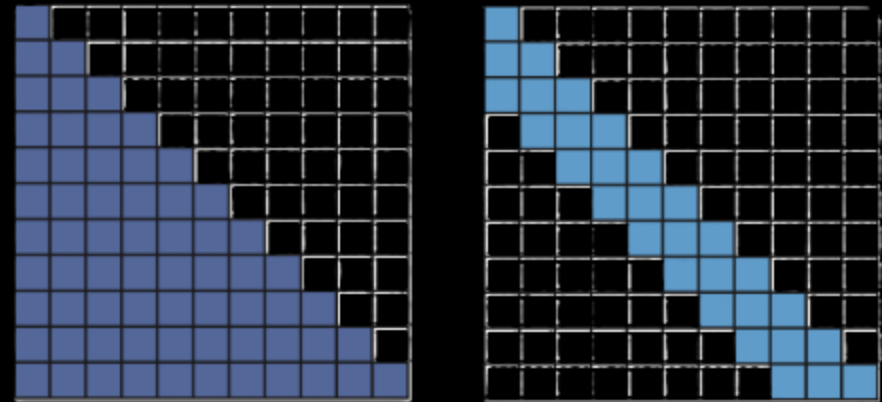
Key Improvements of Krikri over Meltemi

	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	✓	✓



Main architectural diff between Llama-3.1 and Mistral-v0.1

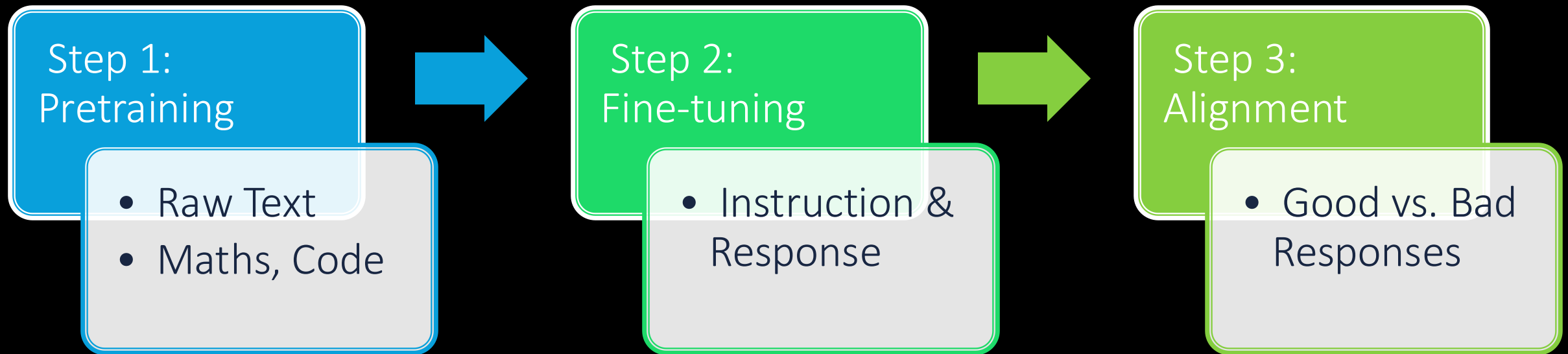
- Full Attention vs Sliding Window Attention
 - Full Attention: Better in long context retrieval and tasks, less efficient computationally
 - SWA: Better for downstream tasks and more efficient computationally



Currently, models employ both in different numbers and strides.



Training Process in a Nutshell



Tokenizer & Embeddings Expansion

Vocabulary Extension

- Extending tokenizers to include specialized Greek tokens.
 - (a) better capture of context and semantics
 - (b) increase inference speed
 - (c) lower training & deployment costs
 - (d) “Increased” context length (less tokens are spent per word)

Llama 3: 19 tokens

Τα μεγάλα γλωσσικά μοντέλα χρειάζονται καλούς tokenizers

GPT 4o: 17 tokens

Τα μεγάλα γλωσσικά μοντέλα χρειάζονται καλούς tokenizers

Krikri: 9 tokens

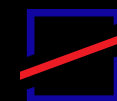
_Τα_μεγάλα_γλωσσικά_μοντέλα_χρειάζονται_καλούς_tokenizers



Vocabulary Extension (cont.)

- The Llama tokenizer is a BPE TikToken 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
- We initialize the BPE vocabulary as the Llama tokenizer vocabulary and continue byte pairing on curated subsets of the training dataset

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



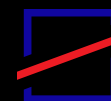
Embeddings expansion

- Untrained embeddings and `lm_head` weights need to be added to the model to match the new tokens
- Initialized by averaging the embeddings (and `lm_head`) of output tokens from the Llama tokenizer, when given a new entry:
 - new token: {"_κανονικός": 100}
 - `krikri_tok("_κανονικός") = [100]`
 - `llama_tok("_κανονικός") = [10, 9, 18, 25]` (e.g.,: ["_κα", "v", "ov", "ικός"])
 - Initialization $\rightarrow \text{emb}[100] = \text{mean}([\text{emb}[10], \text{emb}[9], \text{emb}[18], \text{emb}[25]])$



Embeddings warm-up training

- Light training step to warm start embeddings
- Freeze all model parameters other than:
 - New embeddings
 - New lm_head weights
- ~4% of the pretraining dataset
- Higher learning rate ($3e-4$ cosine annealing down to $3e-5$)



Continual Pretraining & Annealing Phase

Pretraining Data

Νόαμ Τσόμσκι

Λήμμα Συζήτηση

Ανάγνωση

Επεξεργασία

Επεξεργασία κώδικα

Προβολή ιστορικού

Εργαλεία



Το λήμμα δεν περιέχει **πηγές** ή αυτές που περιέχει δεν επαρκούν. Μπορείτε να βοηθήσετε προσθέτοντας την κατάλληλη τεκμηρίωση. Υλικό που είναι ατεκμηρίωτο μπορεί να αμφισβητηθεί και να αφαιρεθεί.

Η σήμανση τοποθετήθηκε στις 27/12/2012.

Ο **Νόαμ Τσόμσκι** (αγγλ.: *Avram Noam Chomsky*, 7 Δεκεμβρίου 1928) είναι **Αμερικανός** καθηγητής στο Τμήμα Γλωσσολογίας και Φιλοσοφίας του Τεχνολογικού Ινστιτούτου της Μασαχουσέτης (MIT)^[32]. Έχει συγγράψει πλήθος βιβλίων και άρθρων, ενώ είχε δώσει και εκτενείς διαλέξεις επάνω σε ένα ευρύτατο φάσμα θεμάτων τα οποία περιλαμβάνουν τη γλωσσολογία, τη φιλοσοφία και την ιστορία της διανόησης. Επίσης, ο Τσόμσκι έχει πλούσιο ιστορικό πολιτικού ακτιβισμού από τη **δεκαετία του '60** κι έπειτα, με πληθώρα βιβλίων τα οποία επικρίνουν κυρίως την εξωτερική πολιτική των ΗΠΑ, ενώ με δήλωσή του τοποθέτησε ιδεολογικά τον εαυτό του στον **αναρχικό** χώρο, αν και κατά καιρούς έχουν σχολιαστεί οι **φιλελεύθερες** καταβολές της πολιτικής του σκέψης^[33]. Σπούδασε στο **Πανεπιστήμιο της Πενσυλβάνια**, απ' όπου αποφοίτησε με πτυχίο φιλοσοφίας το 1949 και με μεταπτυχιακό στη γλωσσολογία το 1951. Η συνεργασία του εκεί με τον διάσημο γλωσσολόγο **Zellig Harris** έθεσε τις βάσεις για τις μετέπειτα θεωρίες του στη γλωσσολογία και την πολιτική του σκέψη.^[34]

Ο Τσόμσκι είναι ο εισηγητής της λεγόμενης «γενετικής-μετασχηματιστικής γραμματικής», κυρίως με το ριζοσπαστικό γλωσσολογικό του σύγγραμμα *Συντακτικές Δομές* του 1957. Η γενετική θεωρία για τη γλώσσα, παρά τις αλλεπάλληλες εξελίξεις και αναθεωρήσεις της, χαρακτηρίζεται από κάποιες σταθερές παραδοχές όπως η ακόλουθη: η ικανότητα του ανθρώπου να παράγει και να κατανοεί ασύλληπτο για τον ανθρώπινο νου αριθμό προτάσεων είναι έμφυτη και μας οδηγεί βάσιμα στην υπόθεση για την ύπαρξη γλωσσικών καθολικών, γενικευμένων δομών και περιορισμών στους οποίους υπακούν όλες οι φυσικές γλώσσες, παρά την παρατηρούμενη τεράστια τυπολογική ποικιλία τους. Η ύπαρξη μιας εγγενούς «Καθολικής Γραμματικής» καθιστά το παιδί ικανό να μαθαίνει τη μητρική του γλώσσα σε ελάχιστο χρονικό διάστημα, παρά την αποσπασματικότητα των δεδομένων τα οποία προσλαμβάνει ο άνθρωπος κατά τα πρώτα χρόνια της ζωής του.

Ο Τσόμσκι αμφισβήτησε τη συμπεριφοριστική θεωρία για την γλωσσική κατάρκτηση και προώθησε ένα τυπικό/**φορμαλιστικό** πρότυπο περιγραφής και ανάλυσης της γλώσσας, ιδιαίτερα στο επίπεδο της σύνταξης. Ο τρόπος με τον οποίο ο Τσόμσκι απέδωσε φορμαλιστικά την γλωσσική ικανότητα εξηγεί επαρκώς τη λεγόμενη "γλωσσική δημιουργικότητα", κεντρική ιδιότητα των φυσικών γλωσσών: με ένα πεπερασμένο αριθμό κανόνων και ένα ορισμένο σύνολο γλωσσικών μονάδων (φωνολογικών και λεξικών), οι άνθρωποι είναι ικανοί να παραγάγουν και να αντιληφθούν άπειρο πλήθος προτάσεων, συμπεριλαμβανομένων προτάσεων που ποτέ ξανά δεν έχουν παραχθεί.

Ο Τσόμσκι αναφέρει ότι τα παιδιά δεν γεννιούνται «*tabula rasa*» (κενό χαρτί) αλλά είναι γενετικά προδιατεθειμένα, ώστε να μπορούν να δομήσουν την απόκτηση της γλώσσας. Σύμφωνα με αυτόν υπάρχει ένας έμφυτος μηχανισμός κατάκτησης της γλώσσας που βοηθά τα παιδιά στο να κατακτήσουν και να αποκτήσουν την γλώσσα τους.

Νόαμ Τσόμσκι



Γενικές πληροφορίες

Όνομα στη μητρική γλώσσα

Νοαμ Τσόμσκι (Αγγλικά)

Γέννηση

7 Δεκεμβρίου 1928^[1]^[2]^[3]

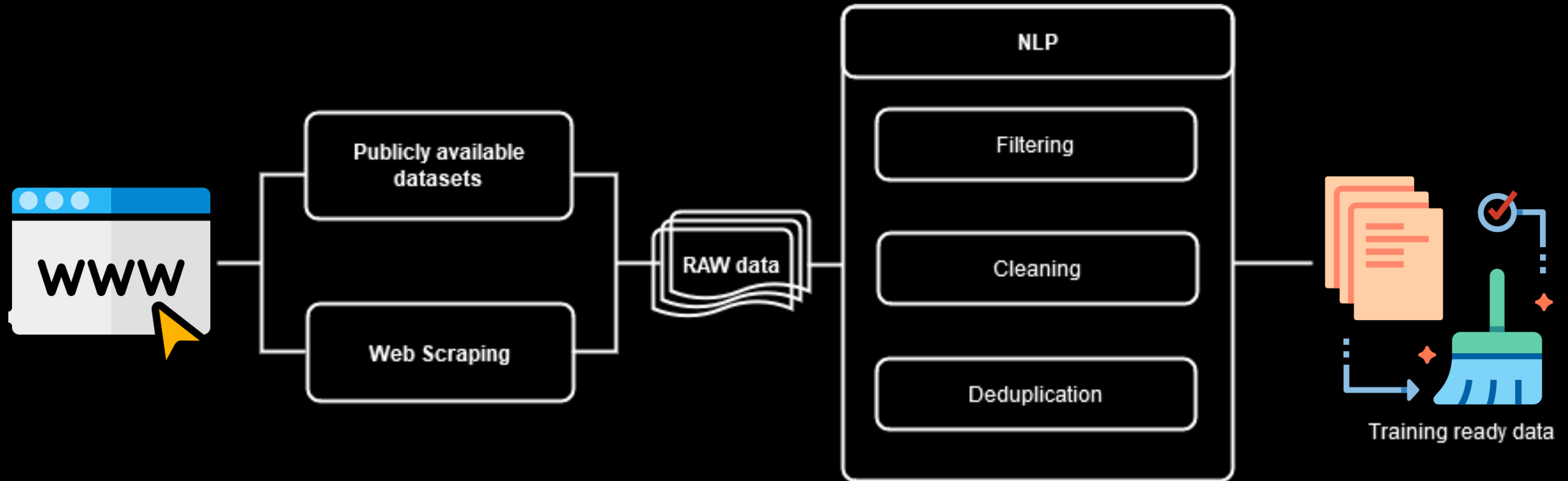
Χοτ Όουκ Λέιν, Φιλαδέλφεια^[4]^[5] ή Φιλαδέλφεια^[6]

Κατοικία

Τούσον
Χοτ Όουκ Λέιν, Φιλαδέλφεια

Athens Natural Language Processing Summer School 2025
September 4-10, 2025 - NCSR "Demokritos"

Data Preprocessing



Data Preprocessing (cont.)

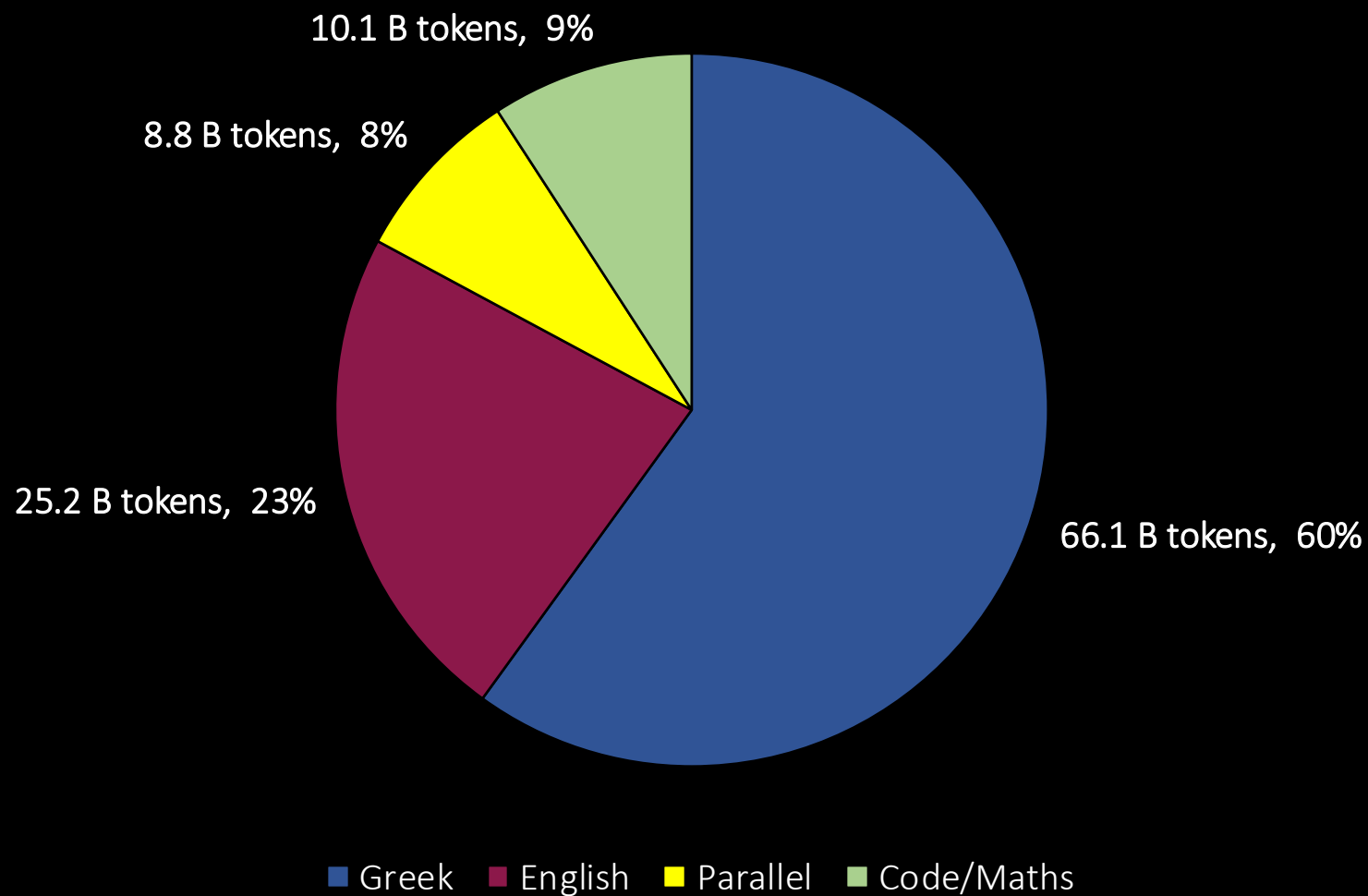
- Text extraction from PDFs, HTMLs, DOCs, etc.
- Conversion in metadata-enriched JSON format
- Pre-processed & filtered using:
 - Rule-based filtering (e.g., mean word length, "lorem ipsum")
 - Scores & Thresholds, such as:
 - Fluency scores with KenLM models
 - Alignment scores for parallel data (LASER, BiCleaner AI)
 - Language identification scores with FastText models
 - Document-level deduplication with MinHash LSH



Composition of pretraining data - Krikri

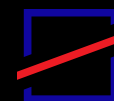
Sources include:

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



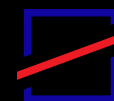
Continual Pretraining

- Largest training stage with 110 B tokens (up-sampled high-quality datasets)
- $1e-4$ max learning rate cosine annealing down to $1e-5$
- 16M tokens global batch size
- Semantically-aware packing into $\sim 128k$ sequence length
- Linear training batch distribution schedule
 - Batches transition:
from mostly English sequences \rightarrow to mostly Greek sequences
 - Code/math are kept stable



Annealing Phase

- Final quality refinement with:
 - 3.5 B tokens from curated & strictly filtered sources
 - 0.2 B tokens from Synthetic Question-Thinking-Answer data
- $1e-5$ max learning rate, linearly annealing down to 0
- 16M tokens global batch size
- Semantically-aware packing into ~128k sequence length
- Attention blocking between batched sequences




Post-training

Supervised Fine-tuning & Alignment

Fine-tuning Data


Teaches model to respond as an assistant, based on a user's prompt.

 **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)**

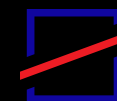
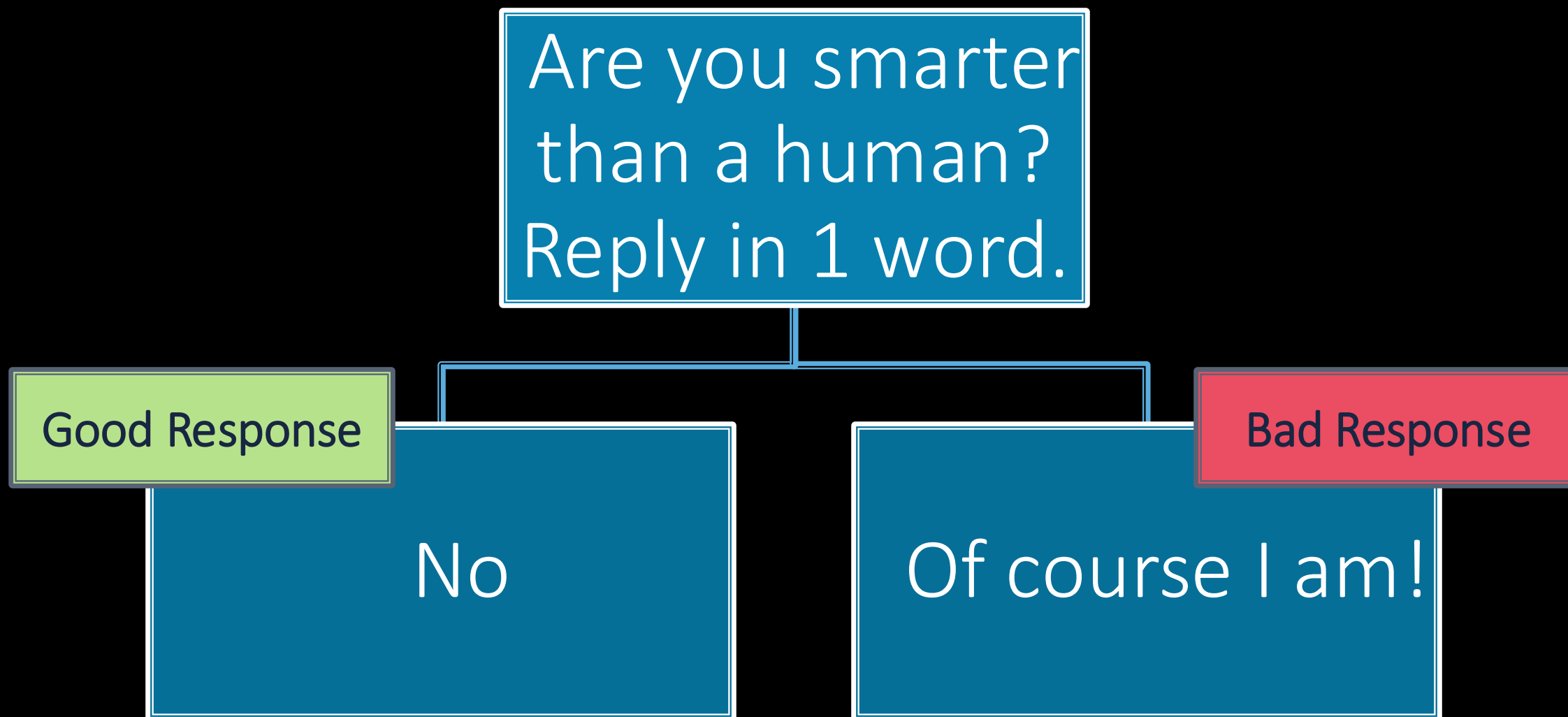
Here are 3 variations of 1-sentence short descriptions of a room:

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?



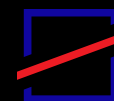
Alignment Data



Synthetic Data Generation

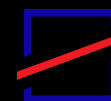
We used various methods to address data scarcity:

- Collect high-quality English SFT & Alignment data
 - **Translate** into Greek & **regenerate** w/ many LLMs (e.g., Gemma 2 27B)
 - 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



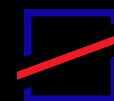
Supervised Fine-Tuning

- **Stage 1:**
 - 856k SFT data
 - 8,192 max sequence length
 - $5e-6$ max learning rate linearly annealing down to $5e-7$
 - 128 samples global batch size
- **Stage 2: Higher quality & thinking data**
 - 638k SFT data
 - 32,768 max sequence length
 - $1e-6$ max learning rate linearly annealing down to $1e-7$
 - 128 samples global batch size



Alignment

- **Length-normalized DPO**
 - Combats bias of aligned models to prefer longer sequences
- **Stage 1: General preference data**
 - 92k preference data
 - 8,192 max sequence length
 - $5e-7$ max learning rate, linearly annealing down to $5e-8$
- **Stage 2: Self-generated preference data (rejection sampling)**
 - 40k preference data from Stage 1 model (6 completions -> RM scoring)
 - 16,384 max sequence length
 - $5e-7$ max learning rate, linearly annealing down to $5e-8$



Preprocessing & Training Infrastructure

- RAM/CPU-heavy nodes:
 - Data collection & preprocessing (Biggest bottleneck -> Deduplication)
 - Tokenizer extension
- Smaller GPU nodes (8 x L40s GPUs = 384 GB):
 - Synthetic data generation
 - Model distillation
 - Evaluation
- 2 x AWS p5en instances (2 x 8 x H200 GPUs = 2,256GB):
 - Embeddings warm-start
 - Continual pre-training
- 1 x AWS p5en instance (8 x H200 GPUs = 1,128 GB):
 - SFT
 - Alignment



Evaluation of LLMs & Benchmark Creation

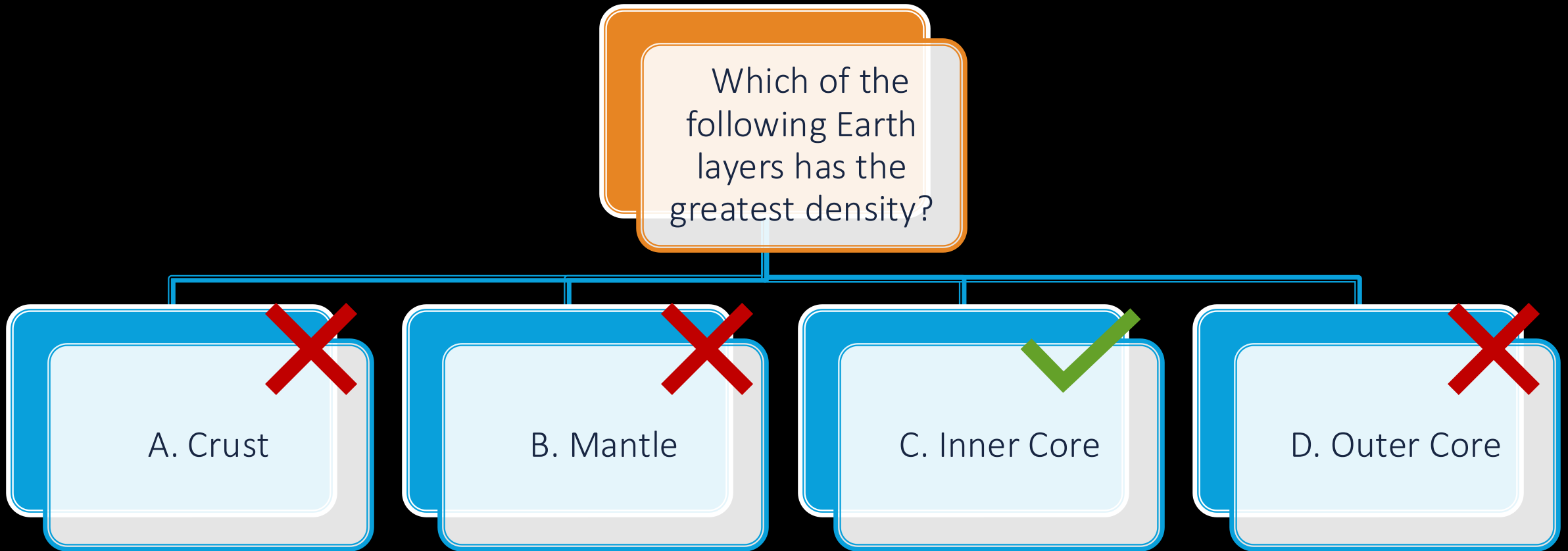
Athens Natural Language Processing Summer School 2025
September 4-10, 2025 - NCSR "Demokritos"

Evaluating LLMs

- Evaluating LLMs involves a combination of methods:
 - Benchmarking on Standard Tasks & Established Test Sets
 - Quantitative Metrics (Accuracy, BLEU, etc.)
 - LLM-as-Judge Methods & Rankings
 - Safety Analysis & Human Evaluation
- **Problem:** Limited Greek benchmarks existed
- **Solution:** We continually extend a complete evaluation suite for Greek
- **How:** Translation, Acquisition, Manual Curation & Annotation



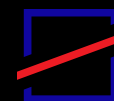
Example: Multiple-Choice Question-Answer



The ILSP LLM evaluation suite for Greek

The evaluation suite comprises 16+ test sets that evaluate:

- Language Understanding and Reasoning
- General Question Answering (Medical, Scientific, Greek Law & History)
- Safety, Jailbreaking, & Truthfulness
- Instruction Following (e.g. write in 3 paragraphs)
- Code & Maths
- General Chat Capabilities
- Translation (even for Ancient Greek)



Recent Benchmarks

- **Greek Physical Commonsense Reasoning (GPCR):**
208 human-annotated data testing physical commonsense reasoning
- **AttaQ Greek:**
1,402 adversarial questions evaluating harmfulness in LLM responses
- **Multiple Choice QA Greek ASEP:**
1,200 multiple choice questions from Greek Civil Personnel Selection exams
- **Greek Medical Multiple Choice QA:**
2,034 multiple choice questions from medical exams of the Hellenic National Academic Recognition and Information Center (DOATAP-ΔΟΑΤΑΠ)



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

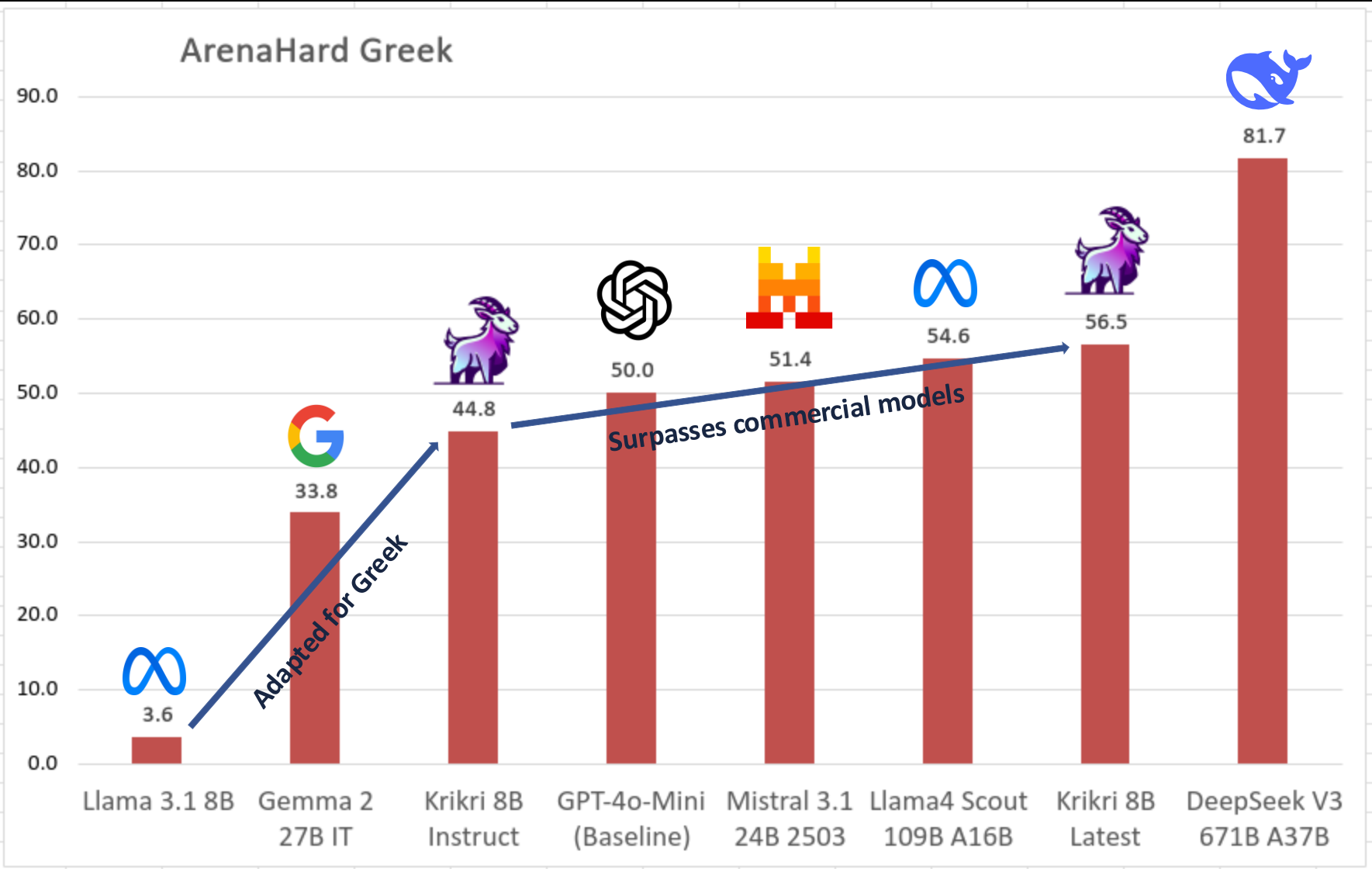
- Outperforms or competes with 27B & 32B models on Greek tasks
- Maintains strong English performance
- Krikri vs Llama 3.1: **+21.7%** on Greek IFEval, **+7.3%** on English IFEval

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

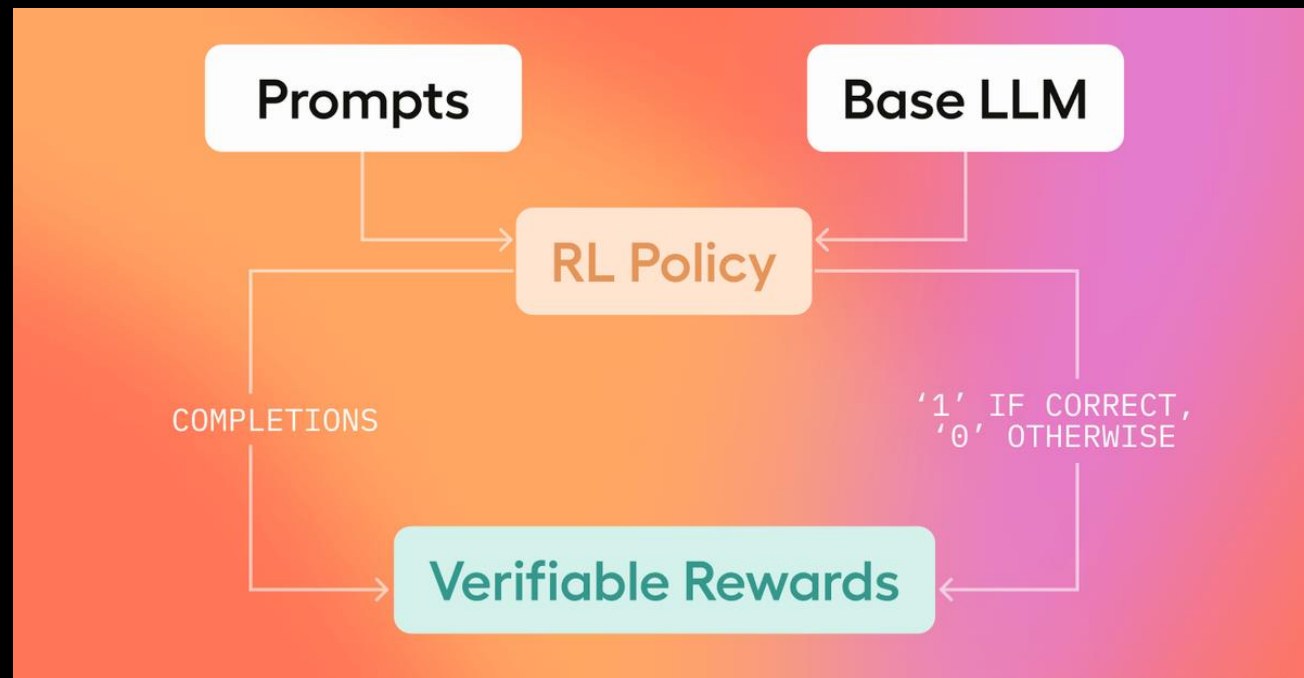
Ranking w/ 500 challenging & diverse prompts (Claude Sonnet 3.7 Judge)



Wrapping up

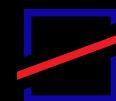
Thinking & Reinforcement Learning

- Krikri was the 1st LLM which natively thinks in Greek
- Targeted capability based on distilled fine-tuning data
- Recent LLMs (e.g., DeepSeek R1) are trained by also using Reinforcement Learning (RL)
- Example: Reinforcement Learning with Verifiable Rewards (RLVR)



Technical Frontiers & Next Steps

- Multimodal extension with **vision & speech**
- Extend training to newer & larger models (Mixture of Experts?)
- Constitutional AI for EU & Greek values and law
- Agentic behavior & deep web research capabilities
- Reinforcement Learning (w/ verifiable and/or proxy rewards)
- Scale up synthetic data generation:
 - Distillation from SOTA models (w/ thinking traces)
 - Better Reward Models and LLMs-as-Judges

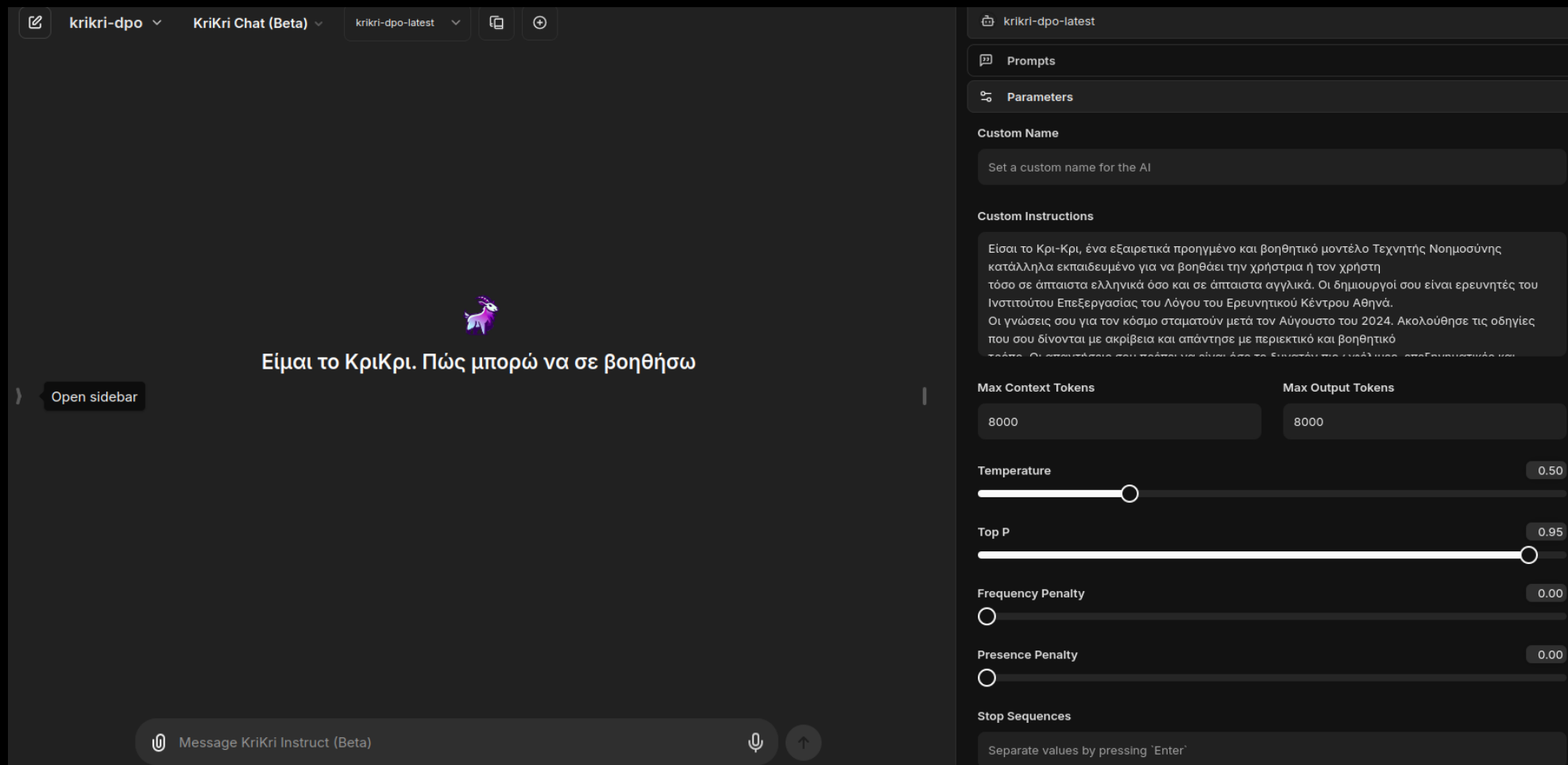


Outcomes

- [Llama-Krikri-8B-Instruct](#)
- 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
- Created evaluation suite with 16+ test sets for Greek LLMs



Chatting with Krikri: chat.ilsp.gr



Acknowledgments



acRE

Open Clouds for Research
Environments



Thank you!
Questions? Comments?

