Krikri: Advancing Open Large Language Models for Greek

Anonymous EMNLP 2025 submission

Abstract

We introduce Llama-Krikri-8B, a cutting-edge Large Language Model tailored for the Greek language, built on Meta's Llama 3.1-8B. Llama-Krikri-8B has been extensively trained on high-quality Greek data to ensure superior adaptation to linguistic nuances. With 8 billion parameters, it offers advanced capabilities while maintaining efficient computational performance. Llama-Krikri-8B supports both Modern Greek and English, and is also equipped to handle polytonic text and Ancient Greek. The chat version of Llama-Krikri-8B features a multi-stage post-training pipeline, utilizing both human and synthetic instruction and preference data, by applying techniques such as MAGPIE. In addition, for evaluation, we propose three novel public benchmarks for Greek. Our evaluation on existing as well as the proposed benchmarks shows notable improvements over comparable Greek and multilingual LLMs in both natural language understanding and generation as well as code generation.

1 Introduction

Recent advancements in AI have been largely driven by the development of large-scale foundation models. Meta's Llama 3 (Grattafiori et al., 2024) fostered a new generation of open models, designed for strong multilingual capabilities, code generation, reasoning, and tool use. With extended context windows, and refined training strategies, models based on Llama 3 have achieved performance comparable to proprietary systems like GPT-4. A critical aspect in this evolution is the development of multilingual and language-specific models, democratizing access to AI technologies and preserving linguistic diversity.

While substantial progress has been made for widely spoken languages, low and medium resource languages remain underrepresented. Greek, in particular, has received limited attention despite its linguistic complexity, rich cultural heritage, and historical significance. Addressing this gap, we present Llama-Krikri-8B, a cutting-edge open Large Language Model tailored for the Greek language. Built on Meta's Llama 3.1-8B architecture, Llama-Krikri has been continually pretrained on a diverse, high-quality Greek corpus. This allows the model to effectively capture the syntactic and semantic nuances of Greek, while retaining the multilingual strengths of the base model. Notably, Llama-Krikri also supports English and is capable of handling polytonic and Ancient Greek texts, addressing not only contemporary but also historical forms of the language.

Compared to Meltemi-7B (Voukoutis et al., 2024), the previous state-of-the-art open Greek LLM built on Mistral 7B (Jiang et al., 2023), Llama-Krikri-8B significantly increases the number of parameters, context length, and training data scale. Additionally, it features an enhanced post-training pipeline using both human and synthetic data. Following the MAGPIE methodology (Xu et al., 2024), we generate high-quality instruction-response pairs via prompting of aligned models, and apply multiple rounds of instruction tuning and alignment using Direct Preference Optimization (DPO) (Rafailov et al., 2024). This pipeline ensures that the model produces helpful, honest, and harmless outputs.

To evaluate Llama-Krikri-8B, we also introduce three novel public benchmarks specifically designed for Greek. These, alongside existing evaluation suites, show that Llama-Krikri outperforms comparable Greek and multilingual LLMs in both natural language understanding and generation, as well as code-related tasks. Moreover, it supports function calling and agentic behaviors, opening new application domains for Greek users. Llama-Krikri-8B is available under the Llama 3.1 Community License Agreement[1].

[1] https://www.llama.com/llama3_1/license/

Our key contributions are:

- We present Llama-Krikri-8B, a state-of-the-art open Greek foundation model based on Llama 3.1, demonstrating strong capabilities in Modern and Ancient Greek, English, and code generation, along with support for function calling and agentic behavior.
- We implement a rigorous post-training pipeline incorporating synthetic instruction tuning via MAGPIE and alignment through DPO.
- We introduce three new benchmarks for evaluating Greek LLMs, covering language understanding, generation, and code tasks.
- We show that Llama-Krikri-8B significantly outperforms existing open Greek and multilingual models across several domains, while supporting advanced features such as function calling.

2 Background and Related Work

Large Language Models (LLMs) have achieved state-of-the-art performance across a wide variety of natural language processing (NLP) tasks. These models are typically trained on massive corpora dominated by English, leading to strong performance in English-language tasks but comparatively weaker capabilities in other languages (Devlin et al., 2019; Brown et al., 2020). As a result, the development of language-specific LLMs has become an active area of research, particularly for under-represented languages.

One prominent strategy for developing such models is continual pretraining, where a pretrained base model is further trained on data in the target language. This approach allows researchers to leverage the general capabilities of large base models while improving performance in specific linguistic domains, without the prohibitive cost of training from scratch (Gururangan et al., 2020).

Several recent projects have successfully applied continual pretraining to adapt existing models to new languages. BgGPT-GEMMA-2-27B-Instruct (Alexandrov et al., 2024) fine-tunes Google's Gemma-2 model (Riviere et al., 2024) for Bulgarian, combining over 100B tokens of Bulgarian and English data and applying techniques such as Branch-and-Merge to mitigate catastrophic forgetting. Similarly, LeoLM (LAION, 2023) and the Sabiá models (Pires et al., 2023) adapt LLaMA and Mistral-based architectures for German and Portuguese, respectively, through targeted continual pretraining.

For Greek, Meltemi-7B represents the first open generative LLM tailored to the language. It was developed by continually pretraining Mistral-7B on a substantial Greek corpus, followed by instruction fine-tuning. While effective, Meltemi's performance is bounded by the size and capabilities of the base model, as well as the limited post-training alignment techniques employed at the time.

Beyond language adaptation, alignment of LLMs to generate helpful, harmless, and honest outputs has become increasingly central. Early approaches such as InstructGPT (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022) rely on multi-stage fine-tuning pipelines involving human feedback or rule-based constraints. More recently, DPO and data synthesis methods like MAGPIE have enabled scalable and effective instruction tuning. MAGPIE, in particular, leverages high-quality prompting of already-aligned models to generate large volumes of instruction-response pairs, demonstrating that synthetic data can rival or surpass human-curated datasets.

These advancements highlight a trend toward bootstrapping high-quality training data using strong base models, especially in low-resource languages. Our work builds on this foundation by employing Llama 3.1 as a base architecture, scaling up parameter count and context length, and applying a more rigorous post-training pipeline, including instruction tuning with MAGPIE-generated data and DPO for alignment.

3 Methodology

Llama-Krikri-8B is based on the Transformer architecture (Vaswani et al., 2023), which has become the de facto standard for large language models. The model inherits its architecture from Meta's Llama 3.1-8B, leveraging the strong foundation in multilingual understanding, code generation, and reasoning provided by Llama 3.1.

Adapting an LLM for the Greek language requires addressing the lack of high-quality Greek data in the massive datasets typically used to train foundation models. Even though Llama 3.1's pretraining corpus comprises trillions of tokens, it struggles to generate coherent Greek text, thus indicating that Greek data is only a tiny fraction of

Subcorpus	Original Tokens (B)	Percentage	Upsampled Tokens (B)	Percentage
Greek	56.7	62.3%	66.1	60.0%
English	21.0	23.1%	25.2	22.9%
Parallel	5.5	6.0%	8.8	8.0%
Math/Code	7.8	8.6%	10.1	9.1%
Total	**91.0**	**100%**	**110.2**	**100%**

Table 1: Composition of the pretraining corpus - original and upsampled

its training data; we should note however that there is limited information on the composition of its pretraining data. Our approach is to perform continual pretraining with Greek and parallel data to infuse the model with Greek knowledge. This training must be done carefully to avoid catastrophic forgetting (Luo et al., 2025) of the base model's prior knowledge in other languages and domains. We tackle this by (a) constructing a high-quality, large-scale Greek corpus, (b) extending and tuning the tokenizer , (c) interleaving Greek training data with datasets containing English, mathematics, code, and parallel data in languages that the model has already been trained on, (d) employing a dataset sampling schedule during training that prefers data closer to the initial llama-3.1 distribution in the beginning, while shifting closer to our true dataset distribution as training continues and (d) re-warming and re-decaying the learning rate (Ibrahim et al., 2024).

Pretraining Data Collection & Cleaning: As a foundation for continual pretraining, we curated a large corpus of texts totalling approximately 91 billion tokens (after filtering and deduplication), which was upsampled to 110 billion tokens for the final pretraining mix. This corpus was constructed with a primary focus on Greek and aiming on retaining and enhancing the original model's capabilities. The distribution included 56.7 billion monolingual Greek tokens (62.3%), 21 billion monolingual English tokens (23.1%), 5.5 billion parallel data tokens (6.0%), and 7.8 billion math and code tokens (8.6%). Table 1 presents the distribution of the pretraining data mix, with more details provided in Appendix A Pretrained data mix.

After corpus collection, we implemented a multi-stage preprocessing and filtering pipeline to ensure a high quality for the pretraining data. Various parts of our filtering methodology have been informed by approaches used in Voukoutis et al. (2024) and large-scale corpus creation efforts such as Zyda (Tokpanov et al., 2024). However, we have adapted these approaches to cater for the peculiarities of the Greek language. We detail the preprocessing pipelines we used in Appendix B Pretraining data cleaning pipelines.

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**

Table 2: Tokenizer statistics for Greek and English

Tokenizer and Embeddings Expansion: The original Llama 3 tokenizer comprises $128,000$ tokens and is inefficient for Greek texts, as it generally performs character-level tokenization for Greek. This was determined, through the approximation of the llama-3.1 tokenizer's fertility (Csaki et al., 2023), a metric of the average tokens per word produced. To determine the efficiency of the original Llama 3 tokenizer and compare with our approach, we conducted tests on diverse Greek and English corpora (each one containing $100,000$ rows and totalling approximately 2M words) and calculated the difference in fertility, as can be seen in Table 2.

In order to develop an optimal tokenizer for Greek which is also efficient in historical dialects of the language, as well as in critical domains (such as legal and scientific texts), we extended the Llama 3 tokenizer with 20,992 new tokens through a multi-stage process which encompasses curating high-quality texts and allocating new tokens across five domains. This process is especially important during model inference, as it significantly reduces the input and output token cost during model use. Furthermore, more compact representations of input text help to improve model performance. We provide detail on the steps for the tokenizer and embeddings expansion in Appendix C, Tokenizer and embeddings expansion process.

Greek embeddings training: To effectively integrate newly introduced Greek tokens into the model, we implemented an initial, targeted training phase for their corresponding embeddings. By preparing the new token representations prior to full-scale pretraining, we mitigate potential disruptions to the existing model parameters.

We initialized the model with Llama 3.1-8B-Base weights, freezing all but the embeddings and output-projection weights for the 20,992 new tokens, allowing their initial training without large gradient updates to the rest of the model. The dataset for this step was comprised of 5B tokens sampled from the main pre-training dataset. This short, several-thousand-step training regimen ensured a smoother integration of the new vocabulary into the model's existing knowledge representation.

Continual Pretraining Process: After embedding training, all parameters were unfrozen, and training continued on the 110B token corpus using a mixed-curriculum strategy. Training alternated between chunks of predominantly Greek text and supporting data (English, parallel, code) in a round-robin fashion. This interleaving improved Greek performance while maintaining/improving English validation perplexity, similar to findings in related work. The strategy involved curriculum learning and experience replay: starting with simpler Greek/more English, progressing to diverse Greek, and mixing in periodic English/code replays. Training was conducted on two machines, each equipped with 8 NVIDIA H200 GPUs, using DeepSpeed Zero 3 and bf16 mixed precision for $\sim$50 days on the 110B token dataset at 128K context length. Training used packed sequences, cosine annealing LR, AdamW optimizer, gradient clipping, and weight decay.

Annealing Phase: Following pretraining, a short annealing pass used a curated 3.5B token dataset of very high-quality texts across all subcorpora. We used within-dataset normalized perplexity, calculated using KenLM (Heafield, 2011), to implement a dataset-aware fluency scoring method for document selection. To boost comprehension and reasoning, a synthetic question-answer dataset (189M tokens) was created by prompting Gemma-2-27B-IT to generate Q&A triplets with reasoning from curated documents. Annealing was tested with and without the synthetic QA data. Performance (Table 3) showed continual pretraining improved Greek (+8.7) but reduced English (-4) vs Llama-3.1. Annealing with curated data gave modest gains. Most notably, adding synthetic QA significantly improved Greek (+2.1 vs continual pretraining) and enhanced English beyond original Llama-3.1 (+0.8). Liger kernels (Hsu et al., 2024) were used for efficiency.

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**

Table 3: Average performance across training stages on Greek and English benchmarks

Instruction Tuning and Alignment: Llama-Krikri-8B-Instruct was created by fine-tuning the base model for instruction following and dialogue. Addressing Greek data scarcity and avoiding translation artifacts, the pipeline combined data synthesis, filtering, two-stage Supervised Fine-Tuning (SFT), and Direct Preference Optimization (DPO).

Data collection, synthesis, & curation involved collecting high-quality English instruction, e.g., Tulu 3 (Lambert et al., 2025) and SmolTalk (Allal et al., 2025), and preference data e.g., UltraFeedback (Cui et al., 2023). Additionally, Greek data was synthesized via translation (with post-editing), regenerating responses using LLMs (Gemma-2-27B-IT), and generating synthetic instructions directly in Greek using the MAGPIE technique. Curated corpora from annealing were reused for synthetic Q&A and dialogues. Data was scored and filtered using the Skyword-Reward-Gemma-2-27B-v0.2 (Liu et al., 2024) reward model, known for Greek accuracy. Rule-based filters ensured formatting and language verification.

SFT was done in two stages ($\sim$856k pairs in Stage 1, $\sim$638k in Stage 2), with progressively higher data quality. Datasets included filtered original English, reward-model-filtered synthetic MAGPIE data (higher scores in Stage 2), translated/post-edited data (Stage 1), regenerated responses (including "thinking" in Stage 2), multi-language translation data, synthetic QA, synthetic multi-turn dialogues, and upsampled manual safety data. Training used linear decay LR, AdamW, and Liger kernels, masking prompt loss. SFT produced a model that followed instructions but needed alignment for helpfulness, precision, and safety.

DPO provided final alignment using $\sim$92k preference triplets. Data included selected/high-scored original preferences, preferences from MAGPIE-synthesized data (using reward model scores), preferences derived from translated data (contrasting regenerated responses or regenerated vs translated reference), and safety preferences. DPO maximized preferred response likelihood while minimizing dispreferred, using length normalization. Training used AdamW (Loshchilov and Hutter, 2017), linear decay LR, and Liger kernels. DPO significantly improved response quality, safety, and helpfulness compared to the SFT-only model. The DPO-tuned model is the final Llama-Krikri-8B-Instruct.

4 Evaluation

In this section, we present evaluation details for Llama-Krikri-8B-Base and Llama-Krikri-8B-Instruct, across six Greek and six English benchmarks. We compare our base model directly with the base model Llama-3.1-8B (Grattafiori et al., 2024) and the previous Greek state-of-the-art model Meltemi-7B-v1.5 (Voukoutis et al., 2024). Additionally, we evaluate our chat model, Llama-Krikri-8B-Instruct on three challenging English benchmarks, as well as three novelly constructed Greek benchmarks which correspond to the English ones.

4.1 Base Model Evaluation: Krikri-8B-Base

We evaluated Llama-Krikri-8B-Base against Llama-3.1-8B and Meltemi-7B-v1.5 in a few-shot setting, consistent with the Open LLM Leaderboard[2].

Greek Benchmarks: The evaluation was carried out on a suite of six Greek-specific benchmarks[3] used in Voukoutis et al. (2024), including machine-translated versions of established English datasets (ARC-Challenge Greek, Truthful QA Greek, HellaSwag Greek, MMLU Greek), the existing Belebele Greek benchmark (Bandarkar et al., 2024), and a novel medical QA benchmark (Medical MCQA).

Results in Table 4 demonstrate substantial improvements for Greek (+10.8%) compared to Llama-3.1-8B. Moreover, we observe that Llama-Krikri-8B-Base surpasses Meltemi-7B-v1.5 with a notable +11.6% average improvement across all benchmarks. On MMLU Greek, Llama-Krikri-8B-Base surpasses Llama-3.1-8B and Meltemi-7B-v1.5 by +9.4% and +10.8% respectively, while on ARC-Challenge Greek, it achieves an accuracy of 49.4%, compared to Llama-3.1-8B's and Meltemi-7B-v1.5's 39.9% and 40.0%, respectively. Similar substantial gains are observed on the Belebele Greek dataset, where Llama-Krikri-8B-Base scores 82.7%, surpassing Meltemi-7B-v1.5 and Llama-3.1-8B by 21.7% and +9.9%, respectively. In the Greek Medical MCQA, Llama-Krikri-8B-Base reaches 53.8%, demonstrating clear advancements over Llama-3.1-8B (+20.4%) in a domain-specific Greek benchmark that was not translated from English.

English Benchmarks For the evaluation of base models on English, we utilized six benchmarks, with five of them being the original versions of those also used for Greek: ARC-Challenge (Clark et al., 2018), Truthful QA (Lin et al., 2022), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and Belebele (Bandarkar et al., 2024). Additionally, the Winogrande (Sakaguchi et al., 2021) test set was used as the sixth benchmark for English. In the results presented in Table 5 we see that our training methodology not only mitigates catastrophic forgetting effectively, but also improves average performance across all English test sets by +0.8%.

4.2 Chat Model Evaluation: Krikri-8B-Instruct

For evaluating the capabilities of Llama-Krikri-8B-Instruct as a conversational assistant, suitable for multi-turn dialogue, instruction-following and complex coding and math queries, we used a suite of benchmarks in both English and Greek. For English, we conducted evaluations across two paths:

- We submitted our model to the Open LLM Leaderboard (Fourrier et al., 2024) which automatically evaluates models on IFEval, BBH, MATH, GPQA, MUSR, and MMLU-Pro using the Eleuther AI Language Model Evaluation Harness (Gao et al., 2021), a unified framework to test generative language models on a large number of different evaluation tasks.

- We used the Arena Hard Auto v0.1 (Li et al., 2024; Chiang et al., 2024), IFEval (Zhou et al., 2023) (strict avg) and MT-Bench (Zheng et al., 2023) benchmarks. Although IFEval was already included in the Open LLM Leaderboard,

[2] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

[3] https://huggingface.co/collections/ilsp/ilsp-greek-evaluation-suite-6827304d5bf8b70d0346b02c

Benchmark	Meltemi-7B-v1.5	Llama-3.1-8B	Krikri-8B-Base
Medical MCQA EL (15-shot)	42.2	33.4	**53.8**
Belebele EL (5-shot)	61.0	72.8	**82.7**
HellaSwag EL (10-shot)	53.8	52.1	**64.6**
ARC-Challenge EL (25-shot)	40.0	39.9	**49.4**
TruthfulQA MC2 EL (0-shot)	49.0	51.1	**54.2**
MMLU EL (5-shot)	41.2	42.6	**52.0**
Average	47.9	48.7	**59.5**

Table 4: Greek benchmark results (accuracy %) for base models.

Benchmark	Meltemi-7B-v1.5	Llama-3.1-8B	Krikri-8B-Base
Winogrande (5-shot)	73.4	**74.6**	72.6
Belebele EN (5-shot)	77.7	71.5	**79.8**
HellaSwag EN (10-shot)	79.6	**82.0**	80.7
ARC-Challenge EN (25-shot)	54.1	**58.5**	57.8
TruthfulQA MC2 EN (0-shot)	40.5	44.2	**44.8**
MMLU EN (5-shot)	56.9	**66.2**	65.1
Average	63.7	66.2	**67.0**

Table 5: English benchmark results (accuracy %) for base models.

we re-implemented it to enable accurate comparison with multiple models. In the evaluation of MT-Bench we used GPT-4o (2024-08-06) as the judge model, while in the evaluation of Arena Hard Auto v0.1 we used the standard approach with GPT-4-0314 as the baseline model (by default scoring 50%) and GPT-4-1106-Preview as the judge model, while also reusing the generations and judgments already computed by the authors.

For Greek, we created three novel evaluation benchmarks by translating three challenging, diverse, and widely used English benchmarks, ensuring high-quality translations through careful post-editing and validation:

- IFEval Greek (strict avg.): a manual translation of 541 prompts from the original Instruction-Following Evaluation benchmark (Zhou et al., 2023), featuring verifiable instructions such as "απάντησε με περισσότερες από 400 λέξεις" (answer with more than 400 words) and "ανάφερε τη λέξη ΤΝ τουλάχιστον 3 φορές" (mention the word AI at least 3 times), designed to assess the model's ability to follow specific instructions.

- MT-Bench Greek: a translated version of the Multi-turn Benchmark (Zheng et al., 2023) containing 80 high-quality, multi-turn conversations across eight diverse categories (e.g., STEM, humanities, roleplay, coding, etc.), carefully post-edited to ensure natural Greek phrasing and cultural appropriateness. MT-Bench is also used to evaluate the function-calling capabilities of LLMs (Chen et al., 2025). The performance of each model is calculated using LLM-as-Judge (Zheng et al., 2023) with GPT-4o (2024-08-06) serving as the scoring model.

- Arena-Hard-Auto Greek: a translated version of Arena-Hard-Auto v0.1, which originates from Chatbot Arena (Chiang et al., 2024) was included in m-ArenaHard (Dang et al., 2024) after translation with Google Translate API v3. We later post-edited using Claude Sonnet 3.5 (Anthropic, 2024) with 10-shot examples to address translation issues, particularly in coding-related prompts where some parts would best be left untranslated, as well as to retain the original style of the prompts, since some of them would be best left vaguely posed as in the original prompt. We used the version of the benchmark with style control methods for Markdown elements[4]. We set GPT-4o-Mini (2024-07-18) as the baseline model (by default 50% score) and GPT-4o (2024-08-06) as the judge model.

As shown in Table 6, Llama-Krikri-8B-Instruct

[4] https://lmsys.org/blog/2024-08-28-style-control/

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

Table 6: Greek and English evaluation results using IFEval and MT-Bench.

demonstrates exceptional performance across both Greek and English benchmarks, substantially outperforming not only its parent model Llama-3.1-8B-Instruct but also other competitive multilingual models in the 7-9B parameter range. It should be noted that the IFEval scores reported in this table reflect our own implementation of the benchmark, which may differ from the Open LLM Leaderboard implementation due to variations in prompt formatting and evaluation criteria. Despite these methodological differences, the relative performance comparisons remain valid within each implementation context.

On IFEval Greek, Llama-Krikri-8B-Instruct achieves a remarkable 67.5% accuracy, surpassing Llama-3.1-8B-Instruct by +21.7% and Meltemi-7B-v1.5 by +34.8%. Notably, our 8B model even outperforms much larger models like Gemma 2 27B IT (+4.3%) and Aya Expanse 32B (+7.2%) on this Greek instruction-following benchmark. As regards the original English IFEval, Llama-Krikri-8B-Instruct scores 82.4%, significantly higher than all other models, including those with 3-4 times more parameters. This dramatic improvement suggests that our data synthesis instruction tuning approach successfully addresses the unique challenges of following instructions in Greek, where naive translations of instruction data often fail to capture language-specific nuances.

For MT-Bench Greek, which evaluates multi-turn conversation quality, Llama-Krikri-8B-Instruct achieves a score of **7.96**, making it the top performer amongst other models in its size class. While larger models like Gemma 2 27B IT (8.23) and Aya Expanse 32B (8.27) achieve slightly higher scores on MT-Bench Greek, the margin is surprisingly small given the substantial difference in model size. On MT-Bench English, Llama-Krikri-8B maintains competitive

Task	Llama-3.1-Instr.	Krikri-8B-Instr.
IFEval	49.22	**60.79**
BBH	**29.38**	29.31
MATH	**15.56**	11.78
GPQA	**8.72**	7.05
MUSR	8.61	**10.46**
MMLU-PRO	**31.09**	25.70
Avg.	23.76	**24.18**

Table 7: Comparative evaluation on English benchmarks from the Open LLM Leaderboard.

performance at 7.21, essentially identical with Llama-3.1-8B-Instruct (-0.04), though understandably lower than the larger Gemma 2 27B IT (-0.79) and Aya Expanse (-0.19). This demonstrates that our instruction tuning approach can achieve a very high conversational performance on Greek conversational benchmarks, while also producing a competitive model for English benchmarks with a much more compact approach (Grattafiori et al., 2024).

As detailed in Table 7, Llama-Krikri-8B-Instruct's official Open LLM Leaderboard submission shows an average score of 24.18% across all tests, slightly surpassing the 23.76% of Llama-3.1-8B-Instruct. The model shows particularly impressive gains on IFEval implementation (60.79% vs. 49.22%) and MUSR (10.46% vs. 8.61%), while closely matching performance on the Big Bench Hard (BBH) benchmark (29.31% vs. 29.38%). Although Llama-Krikri performs slightly below Meta-Llama-3.1-8B-Instruct in the MMLU-PRO category (25.70% vs. 31.09%), the overall performance indicates successful retention of English capabilities during the Greek-focused continual pretraining.

The results from our Arena Hard evaluations, presented in Table 8, reveal that, in the 8B parameter range, Llama-Krikri-8B-Instruct significantly outperforms its competitors, achieving a 31.8%

Model	ArenaHard EL	ArenaHard EN
Aya Expanse 8B	23.8	–
Llama 3.1 8B Instr.	4.0	19.7
Krikri 8B Instr.	**31.8**	**35.1**
Aya Expanse 32B	40.1	45.1
Gemma 2 27B IT	32.2	49.6
Llama 3.1 70B Instr.	27.4	53.9
GPT 4o Mini	50.0	65.0

Table 8: Arena Hard evaluation results (% win rate) for Greek and English.

win rate on Arena Hard Greek compared to Aya Expanse 8B's 23.8% and Llama 3.1 8B Instruct's 4.0% (+27.8% improvement). This demonstrates the effectiveness of our Greek-focused training approach. Even more impressively, Llama-Krikri-8B-Instruct achieves a 35.1% win rate on Arena Hard English, substantially outperforming the original Llama-3.1-8B-Instruct (19.7%) by +16.2%, despite our focus on Greek capabilities. While Aya Expanse 32B leads on Arena Hard Greek with 40.1%, our 8B model is on par with Gemma 2 27B IT (31.8% vs. 32.2%) and outperforms the 8.75 times larger Llama-3.1-70B-Instruct (27.4%) by +4.4% on the Greek evaluation data. On the original English Arena Hard, the larger models generally perform better, although it should be noted that Llama-Krikri-8B-Instruct outperforms Llama-3.1-8B-Instruct by +15.4% (35.1% vs. 19.7%).

Please note that while all models trail behind GPT-4o-Mini (used as baseline on the Greek Arena Hard), recent research (Li et al., 2025) has shown that judge models are biased towards student models, i.e., models finetuned on distilled data from the stronger & larger teacher model which also acts as a judge. While details on the post-training data of GPT-4o-Mini are undisclosed, it would be very reasonable to assume that it has been trained -at least partly- with GPT-4 and GPT-4o serving as teacher models and, therefore, that the judges that we utilized are biased towards it compared to all other evaluated models.

This performance comparison with much larger models highlights the efficiency of our approach since Llama-Krikri-8B-Instruct achieves comparable or even superior performance on Greek benchmarks compared to models with 3-4x more parameters, while maintaining strong English capabilities. This efficiency is particularly important for deployment scenarios where computational resources may be limited, demonstrating that a carefully trained smaller model can rival much larger ones for specific languages.

Apart from the comparative evaluations mentioned above, we have performed zeroshot MT experiments on an Ancient-Modern Greek (grc↔ell) translation dataset[5] that includes 100 sentences of Ancient Greek texts manually translated into Modern Greek. Using Llama-Krikri-8B-Instruct we have observed a 54.66 BLEU score for the Ancient to Modern Greek (grc→ell) translation direction, with the reverse direction (ell→grc) being more challenging (20.41 BLEU).

5 Discussion and Conclusions

In this paper, we presented Llama-Krikri-8B, a new LLM that exhibits significant skills in understanding and generating Greek, while also showing highly accurate handling of text in English and historical Greek dialects. We achieved this by developing an efficient tokenizer that exhibits a low token/words fertility for Greek and by further training Llama 3.1-8B using a carefully constructed dataset that covered a wide variety of domains. In evaluation experiments on a benchmark suite comprising Greek and English datasets, we have observed that Llama-Krikri-8B performs significantly better in Greek (+10.8%) compared to its base model, while also showing gains in English (+0.8%). We then created Llama-Krikri-8B-Instruct, a version designed for following instructions and engaging in helpful conversations. This involved a multi-step process that comprised synthetic data generation in a multitude of domains, fine-tuning the model and then aligning it with human preferences. Evaluations revealed that Llama-Krikri-8B-Instruct significantly outperformed Llama-3.1-8B-Instruct in both Greek (+21.7%) and English (+7.3%) IFEval. Our model also demonstrated highly competitive chat abilities in both languages across several benchmarks.

6 Limitations

The quality and accessibility of Greek datasets are critical to the development of Krikri. Greek open-source corpora are becoming more numerous, but they might not be as large or varied as datasets for more extensively spoken languages, like English. This may result in biases in the model's understanding of Greek, especially with regard to regional variances, dialects, and specialized fields like technical fields, law, or medicine.

[5] https://huggingface.co/datasets/ilsp/ancient-modern_greek_translations

As an 8B parameter model, our model shows a fairly high level of Greek fluency, but it is less effective than larger-class and commercial models at reasoning and instruction following, and is more likely to experience hallucinations.

In the future, our evaluation benchmarks should include more original Greek LLM datasets that are not the result of machine translation and post-editing. These datasets will help minimize the effect of machine translation on evaluation results and also better reflect the target language and culture.

7 Risks and ethical considerations

To mitigate potential risks, we took several steps to ensure the data used for training did not contain personally identifiable information, offensive, or otherwise inappropriate content. We sourced data from publicly available, licensed, or open-access datasets, ensuring compliance with their respective policies and any flagged data points were excluded. We did not collect data from private communications or data sources that could contain personally identifiable information. We have also given special care to align our model's responses with safety guidelines followed by manual reviews.

However, we have not performed a systematic evaluation against LLM risks including risks related to discrimination, hate speech and exclusion; information hazards; and misinformation harms (Weidinger et al., 2022). As this was due to lack of relevant evaluation material in Greek, we want to contribute towards improving this situation in the future.

We recognize that these measures are not a substitute for more thorough evaluation protocols. Moving forward, we aim to contribute towards addressing these limitations by promoting the development of Greek-language evaluation resources for LLM risks. This will enable more robust and contextually appropriate assessments of ethical risks in future models.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Anton Alexandrov, Veselin Raychev, Dimitar I. Dimitrov, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Bggpt 1.0: Extending english-centric llms to other languages. *Preprint*, arXiv:2412.10893.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 749–775.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

A Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, page 21.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*. OpenReview.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Yi-Chang Chen, Po-Chun Hsu, Chan-Jan Hsu, and Dashan Shiu. 2025. Enhancing function-calling capabilities in LLMs: Strategies for prompt formats, data integration, and multilingual translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 99–111, Albuquerque, New Mexico. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *International Conference on Machine Learning*, pages 8359–8388. PMLR.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. *arXiv preprint arXiv:2311.05741*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darundefinedis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Lang. Resour. Eval.*, 57(1):415–448.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Maria Gavriilidou, Stelios Piperidis, Dimitrios Galanis, Kanella Pouli, Penny Labropoulou, Juli Bakagianni, Iro Tsiouli, Miltos Deligiannis, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, and Katerina Gkirtzou. 2023. The CLARIN:EL infrastructure: Platform, Portal, K-Centre. In *Selected papers from the CLARIN Annual Conference 2023*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. 2024. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

LAION. 2023. LeoLM: Igniting German-Language LLM Research. `https://laion.ai/blog/leo-lm/`. Accessed: (12 July 2024).

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets*. Cambridge university press.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025. Preference leakage: A contamination problem in llm-as-a-judge. *Preprint*, arXiv:2502.01534.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Andrea Lösch, Valérie Mapelli, Khalid Choukri, Maria Giagkou, Stelios Piperidis, Prokopis Prokopidis, Vassilis Papavassiliou, Miltos Deligiannis, Aivars Berzins, Andrejs Vasiljevs, et al. 2021. Collection and Curation of Language Data within the European Language Resource Coordination (ELRC). In *Qurator*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on greek legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 63–75, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis, and Stelios Piperidis. 2018. The ILSP/ARC submission to the WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 928–933, Belgium, Brussels. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Richard J Roberts. 2001. Pubmed central: The genbank of the published literature.

Dimitrios Roussis and Vassilis Papavassiliou. 2022. The ARC-NKUA submission for the English-Ukrainian general machine translation shared task at WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 358–365, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022a. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France. European Language Resources Association.

Dimitrios Roussis, Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis, and Stelios Piperidis. 2022b. Constructing parallel corpora from COVID-19 news using MediSys metadata. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1068–1072, Marseille, France. European Language Resources Association.

Dimitris Roussis, Sokratis Sofianopoulos, and Stelios Piperidis. 2024. Enhancing scientific discourse: Machine translation for the scientific domain. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 275–285.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Liping Tang, Nikhil Ranjan, Omkar Pangarkar, Xuezhi Liang, Zhen Wang, Li An, Bhaskar Rao, Linghao Jin, Huijuan Wang, Zhoujun Cheng, Suqi Sun, Cun Mu, Victor Miller, Xuezhe Ma, Yue Peng, Zhengzhong Liu, and Eric P. Xing. 2024. Txt360: A top-quality llm pre-training dataset requires the perfect blend.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Yury Tokpanov, Beren Millidge, Paolo Glorioso, Jonathan Pilault, Adam Ibrahim, James Whittington, and Quentin Anthony. 2024. Zyda: A 1.3 T

Dataset for Open Language Modeling. *arXiv preprint arXiv:2406.01981*.

Santosh Tyss, Rashid Haddad, and Matthias Grabmair. 2024. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The first open large language model for greek. *Preprint*, arXiv:2407.20743.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yifan Zhang, Yifan Luo, Yang Yuan, and Andrew Chi-Chih Yao. 2024. Automathtext: Autonomous data selection with language models for mathematical texts. *arXiv preprint arXiv:2402.07625*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Pretrained data mix

The pretraining data mix (Table 1) contains the following four components:

1. **Greek Texts (56.7B tokens):** The Greek part of the dataset was sourced from publicly available resources spanning a wide range of domains and sources: Wikipedia, ELRC-SHARE (Lösch et al., 2021), EUR-LEX & MultiEUR-LEX (Chalkidis et al., 2019, 2021), MaCoCu (Bañón et al., 2022), CLARIN-EL (Gavriilidou et al., 2023), EMEA[6], parliamentary proceedings (Erjavec et al., 2022), [7], governmental and legal documents from the Greek Government Gazette via the National Printing House[8], the Permanent Greek Legislation Code – Raptarchis dataset[9] (Papaloukas et al., 2021), Greek School Books[10], the Kallipos initiative of Greek open academic textbooks[11], full texts from publicly available articles, theses, and dissertations from academic repositories and the National Documentation Center[12], as well as pre-filtered resources originally compiled from the web, such as CulturaX (Nguyen et al., 2023) and CulturaY[13]. In addition to Modern Greek, we incorporated a significant amount of Ancient Greek texts into our training corpus from Wikisource, school books, web pages, and Project Gutenberg[14], which provides freely available Ancient Greek texts, including classical literature and historical documents. By including

[6] https://www.ema.europa.eu/
[7] https://www.gutenberg.org/
[8] https://et.gr/
[9] https://huggingface.co/datasets/AI-team-UoA/greek_legal_code
[10] https://ebooks.edu.gr/ebooks/
[11] https://kallipos.gr/en/homepage/
[12] https://www.ekt.gr/en
[13] https://huggingface.co/datasets/ontocord/CulturaY
[14] https://www.gutenberg.org/

Ancient Greek data, we ensured that Llama-Krikri-8B is able process polytonic Greek and engage with historical texts effectively. This enhances the model's utility for classical studies, historical research, and philological applications.

2. **English Texts (21B tokens):** A subset of high-quality English data was mixed into the training corpus. This subset ensures that the model is continually trained on English data, and is drawn from sources that were also used for the Greek data, such as Wikipedia, Wikisource, Project Gutenberg (post-1900), EUR-LEX, EMEA, Greek academic repositories, etc. We also utilized additional English texts originating from abstracts and full texts of academic records found on multiple scientific repositories (Roussis et al., 2022a, 2024), ECtHR-PCR (Tyss et al., 2024), and pre-filtered datasets from TxT360 (Tang et al., 2024), like ArXiv, S2ORC (Lo et al., 2020), and PubMed Central (Roberts, 2001). By incorporating diverse and high-quality English texts, we mitigate the risk of catastrophic forgetting.

3. **Parallel Data (5.5B tokens):** We compiled a diverse parallel corpus with language pairs covering multiple languages: to Greek, English, French, Portuguese, German, Spanish, and Italian. The decision to add parallel data which covers other European languages (i.e., German, French, Italian, Portuguese, and Spanish) is informed from the languages that have been included in the multilingual instruction tuning of the original Llama-3.1. We utilized resources such as SciPar (Roussis et al., 2022a), MediSys (Roussis et al., 2022b), MultiEUR-LEX (Chalkidis et al., 2021), Europarl, TED Talk transcripts (Qi et al., 2018), and other sources with sentence pairs such as ELRC-SHARE (Lösch et al., 2021) & OPUS (Tiedemann, 2012). Our data include parallel documents and sentence pairs randomly sampled for each translation direction, e.g., EN-EL/EL-EN and EN-DE/DE-EN, as well as augmented training examples with concatenated parallel content across multiple languages (e.g., a Greek text followed by its English, German, and Spanish translations with appropriate prompt templates). The addition of these documents has a twofold effect. It has been shown that parallel data boosts translation performance (Alves et al., 2024; Martins et al., 2024), while limited empirical evidence indicates that pretrained LLMs process multilingual queries by first translating the content into English, utilizing their English knowledge to answer the query and then translate the answer back to the original language(Zhao et al., 2024).

4. **Code and Math (7.8B tokens):** We also integrated datasets containing text with code and mathematics, leveraging Stack Overflow[15], Python-Edu which is a subset of the SmolLM corpus (Allal et al., 2025) originating from The Stack V2 dataset (Lozhkov et al., 2024) and having been scored with an educational code classifier, and the AutoMathText dataset (Zhang et al., 2024), which is a collection of math-related documents originating from web data, papers on arXiv, and code/notebooks on GitHub. AutoMathText has undergone an automatic selection process using Qwen-72B (Bai et al., 2023) for relevancy to the mathematical domain and the educational value of each document. Code and Mathematics data, although not specific to Greek, were included to preserve and enhance the model's ability to handle coding tasks, math problems and formal language. Maintaining these capabilities broadens the utility of Llama-Krikri beyond pure language tasks.

B Pretraining data cleaning pipelines

Our filtering processes began with format standardization in order to facilitate uniform processing across multiple heterogeneous datasets. We converted all textual content from various formats (e.g., PDF, HTML, plain text, etc.) into JSONL containing both the document text and relevant metadata such as identified language, word count, and source information (including source URLs).

For PDF documents such as academic records and laws, we implemented a specialized pipeline which integrated Marker[16] for extraction and conversion into Markdown files, as it exhibits strong performance for Greek texts. Subsequently, the pipeline included language identification using

[15] https://huggingface.co/datasets/code-rag-bench/stackoverflow-posts

[16] https://github.com/VikParuchuri/marker

FastText (Joulin et al., 2017), removal of markdown artifacts, and removal of lines with characters outside Unicode ranges for Greek, Latin, and other common and scientific symbols. Furthermore, we utilized document structure metrics (Marker also extracts various structural metadata) as quality indicators, such as the ratio of tables to pages and the fraction of removed lines in disallowed scripts.

Our main filtering pipeline used sequential rule-based and statistical filters to remove outlier documents across all data sources. First, we implemented URL-based filtering by removing content from several blacklisted domains known to contain low-quality or problematic content. This was particularly effective for web-crawled datasets like CulturaX (Nguyen et al., 2023) where source metadata was available. We then applied a set of minimal content-quality filters:

- Removal of documents containing multiple instances of profane or inappropriate terms from a curated list of Greek bad words
- Removal of short documents based on character and word counts
- Removal of documents containing multiple substrings like "lorem ipsum" which are indicative of content with low educational value
- Removal of documents containing extremely long words (>60 characters)
- Removal of documents with mean word length outside specified values.
- Removal of documents with a high fraction of non-alphanumeric characters.

Parallel datasets were filtered using a different pipeline featuring various steps from previous work (Papavassiliou et al., 2018; Roussis and Papavassiliou, 2022; Roussis et al., 2024) which include: (a) rule-based filters, such as length ratio, language identification verification, and (b) model-based alignment quality scores using tools like LASER (Artetxe and Schwenk, 2018, 2019), BiCleaner AI (Zaragoza-Bernabeu et al., 2022), and CometKiwi (Rei et al., 2022).

Additionally, in order to mitigate privacy concerns and protect sensitive information, we systematically identified and anonymized personally identifiable information (PII) with the use of regular expressions. In particular, we aimed to detect and replace e-mail addresses with a generic placeholder ("email@example.gr") and mask IP addresses (replacing them with 0.0.0.0).

Finally, for Greek, English, and Mathematics/Code datasets we implemented intra-dataset deduplication, as well as cross-dataset deduplication. We utilized MinHashLSH near-deduplication (Broder, 1997; Leskovec et al., 2020) with 5-gram subsets, a MinHash signature of 128, and a Jaccard similarity threshold of 0.8, following parameter choices similar to those used in other works (Nguyen et al., 2023; Voukoutis et al., 2024). Regarding the deduplication of parallel datasets, we followed a different approach. All sentence pairs were normalized and cleaned, by converting them to lowercase and removing digits, punctuation. Pairs were then deduplicated based on the existance of either the source or target within the same dataset, thus ensuring that no sentence can be found multiple times in each parallel dataset (Roussis and Papavassiliou, 2022; Roussis et al., 2024). It should be noted that deduplication has consistently been shown to lead to higher performance, reduced training costs, as well as reduced model memorization; thus indirectly protecting sensitive information (Lee et al., 2022; Carlini et al., 2023; Grattafiori et al., 2024; Albalak et al., 2024). However, as we mentioned earlier, global deduplication may also remove documents of high quality and actually hurt performance (Tang et al., 2024; Penedo et al., 2024). For this reason, we decided to upsample datasets of specific sources with important content. Table 1 summarizes the composition of the filtered and deduplicated pretraining corpus. In total, our collected dataset comprises roughly 91B tokens, of which 62.3% is Greek text. For the final training curriculum, we upsampled parts of the corpus to effectively train on an equivalent of 110B tokens. Upsampling was used to give higher relative importance to certain underrepresented but valuable segments and it also leads to higher memorization of important content (Carlini et al., 2023; Tang et al., 2024). For example, we assigned a slightly higher weight to datasets with long-context documents, Wikipedia-like sources, dialogue data, multi-parallel documents, and to certain important domains, such as legal, scientific, and medical. The decision to include a significant amount of English and parallel data (23.1% and 6% of tokens, respectively) was guided by prior work (Voukoutis et al., 2024) showing that mixed-language training can help retain the base model's general knowledge and prevent catastrophic forgetting.

C Tokenizer and embeddings expansion process

The tokenizer and embeddings expansion process involved the following steps:

- **Data Acquisition:** We acquired data by collecting sentences from high-quality sources of our pretraining mix in five domains:
  1. **General domain** which reuses a sampled portion of the data used to train the tokenizer of Meltemi (Voukoutis et al., 2024) and covers diverse domains,
  2. **Legal domain** which uses legal texts extracted from the Greek Government Gazette and are available via the National Printing House[17], as well as the Permanent Greek Legislation Code – Raptarchis dataset[18] (Papaloukas et al., 2021),
  3. **Scientific domain** which uses publicly available articles, theses, and dissertations found in the National Documentation Center[19],
  4. **Literature domain** from public-domain books from Wikisource[20] and Project Gutenberg[21] which contain literature, poetry, and other original writings across various variants of Greek (e.g., Koine Greek, Medieval Greek, Modern Greek, etc.),
  5. **Ancient Greek** which contains texts only in Ancient Greek sourced from Greek school books and various publicly available corpora.

- **Filtering and Preprocessing:** Each dataset underwent sequential processing and filtering including language identification verification with FastText (Joulin et al., 2017), application of regular expressions to remove URLs and other anomalies, symbol-to-word ratio filtering to remove outliers, and NFC normalization. We then performed sentence-level exact deduplication within each individual dataset.

  To ensure text quality, we applied fluency scoring using Monocleaner (Sánchez-Cartagena et al., 2018) which leverages a 7-gram KenLM model for Greek, and setting a score threshold of 0.3 for non-polytonic text.

- **Creation of Training and Test Sets:** For the tokenizer training and test sets creation, we sampled 50% of the sentences from each source and divided it into training and testing splits (80%–20%).

- **Domain-specific Token Allocation:** New tokens were added sequentially for each domain until tokenizer fertility for this domain remained relatively stable, with most of the tokens being allocated to the General domain. This approach ensured that common Modern Greek patters receive the largest coverage, while specialized terminology and older Greek variants are adequately represented.

Table 2 presents the fertility of several tokenizers on the original Greek and English corpora. Note that these test sets have not been used anywhere in the domain-specific multi-stage training process and could be considered as out-of-domain for the Llama-Krikri-8B tokenizer. We evaluated the Llama-3.1-8B tokenizer (128,000 vocabulary size) and our custom-trained Llama-Krikri-8B tokenizer (149,248 vocabulary size). We observe that the Llama-3.1-8B tokenizer exhibits a fertility of 2.73 for Greek and 1.33 for English. Our Llama-Krikri-8B tokenizer demonstrates a significantly lower fertility of 1.65 for Greek, while maintaining the same low fertility of 1.33 for English as the base Llama-3.1-8B tokenizer. This indicates that our Llama-Krikri-8B tokenizer is more efficient for Greek texts compared to the standard Llama-3.1-8B tokenizer. The table also includes the Mistral and the Meltemi-7B Greek tokenizer for comparison.

Furthermore, in Table 9, we list the token allocation per domain, as well as the tokenizer fertilities of Llama-3.1-8B and Llama-Krikri-8B for each of the five domains for which we created test sets during the tokenizer extention process. We can observe that tokenizer fertility has dropped significantly across all domains, with Ancient Greek benefiting from the highest reduction (-1.62) with just 492 added tokens, whereas fertility for Legal texts has almost halved (1.54 from 2.82) after the integration of a significant number of tokens (4,000).

[17] https://et.gr/
[18] https://huggingface.co/datasets/AI-team-UoA/greek_legal_code
[19] https://www.ekt.gr/en
[20] https://el.wikisource.org/
[21] https://gutenberg.org/

Domain	Added Tokens	Llama-3.1-8B	Llama-Krikri-8B	Δ Fertility
General	15,000	2.65	**1.59**	-1.06
Legal	4,000	2.82	**1.54**	-1.28
Scientific	1,000	2.91	**1.73**	-1.18
Literature	500	2.90	**1.89**	-1.01
Ancient Greek	492	3.77	**2.15**	-1.62
Total	**20,992**	–	–	–

Table 9: Domain-specific token allocation and fertility comparison

Following the development of the Llama-Krikri-8B tokenizer with 149,248 tokens, containing the initial Llama 3 reserved tokens to reach a vocabulary size divisible by 256 for computational efficiency, we expanded the embeddings and lm head layer of our Llama 3 based model to match this vocabulary size.

The newly initialized embeddings and lm head weights for each new token were initialized by averaging the embeddings (or lm head weights respectively) associated with the tokens retrieved by the initial Llama-3.1 tokenizer when fed with that token. This provided us with a smoother initial Greek embeddings and lm head training phase of the frozen model.

D Examples of queries and responses

Ex. 1: Programming: Python Fibonacci function

User: Γράψε μια συνάρτηση Python που υπολογίζει την ακολουθία fibonacci.
Krikri: Ακολουθεί η υλοποίηση της αναδρομικής συνάρτησης Python για τον υπολογισμό των αριθμών Fibonacci:

```python
def fibonacci(n):
    """
    Υπολογίζει τον n-οστό αριθμό Fibonacci.

    Args:
        n (int): θετικός ακέραιος. n >= 0 (0-indexed προσέγγιση)

    Returns:
        int: ο n-οστός αριθμός Fibonacci

    Raises:
        ValueError: αν n είναι αρνητικός αριθμός
    """
    # Έλεγχος εγκυρότητας εισόδου
    if not isinstance(n, int) or n < 0:
        raise ValueError("Η είσοδος πρέπει να είναι μη αρνητικός ακέραιος")

    # Βάση αναδρομής:
    # F(0) = 0, F(1) = 1 (κλασικοί ορισμοί)
    if n in (0, 1):
        return n

    # Αναδρομικός τύπος: F(n) = F(n-1) + F(n-2)
    return fibonacci(n-1) + fibonacci(n-2)
```

Παράδειγμα χρήσης**:**

```python
# Υπολογισμός μερικών όρων της
# ακολουθίας Fibonacci
for i in range(10):
    print(f"F({i}) = {fibonacci(i)}")

# Εκτύπωση αποτελεσμάτων
# F(0) = 0
# F(1) = 1
# F(2) = 1
# F(3) = 2
# F(4) = 3
# F(5) = 5
# F(6) = 8
# F(7) = 13
# F(8) = 21
# F(9) = 34
```

Ex. 2: Translation: Ancient to Modern Greek

User: Μετάφρασέ μου στα Νέα ελληνικά το κείμενο:
‘Οἱ μὲν δὴ Μήλιοι τοσαῦτα ἀπεκρίναντο· οἱ δὲ Ἀθηναῖοι διαλυόμενοι ἤδη ἐκ τῶν λόγων ἔφασαν «ἀλλ’ οὖν μόνοι γε ἀπὸ τούτων τῶν βουλευμάτων, ὡς ἡμῖν δοκεῖτε, τὰ μὲν μέλλοντα τῶν ὁρωμένων σαφέστερα κρίνετε, τὰ δὲ ἀφανῆ τῷ βούλεσθαι ὡς γιγνόμενα ἤδη θεᾶσθε, καὶ Λακεδαιμονίοις καὶ τύχῃ καὶ ἐλπίσι πλεῖστον δὴ παραβεβλημένοι καὶ πιστεύσαντες πλεῖστον καὶ σφαλήσεσθε.»’
Krikri: Ακολουθεί η μετάφραση του κειμένου στα Νέα Ελληνικά:
Οι Μήλιοι απάντησαν με αυτά τα λόγια. Οι Αθηναίοι, καθώς αποχωρούσαν από τη συζήτηση, είπαν: ‘Όμως εσείς είστε οι μόνοι που από αυτές τις συζητήσεις συμπεραίνετε σωστά τόσο τα ορατά όσο και τα αόρατα, σαν να έχουν ήδη συμβεί. Είστε υπερβολικά αισιόδοξοι και έχετε εναποθέσει τις ελπίδες σας στους Λακεδαιμονίους, στην τύχη και στις προσδοκίες σας. Και όσο περισσότερο έχετε πιστέψει σε αυτά, τόσο περισσότερο θα απογοητευτείτε.’

Ex. 3: Translation to JSON Output: Greek to Multiple Languages

User: Μετάφρασέ μου το παρακάτω κείμενο από τα νέα ελληνικά στα αγγλικά, στα ισπανικά και στα γαλλικά. Δώσε μου τις μεταφράσεις σε JSON.

Σύμφωνα με την οδηγία 2009/28/ΕΚ του Ευρωπαϊκού Κοινοβουλίου, ως ενέργεια από ανανεώσιμες μη ορυκτές πηγές θεωρείται η αιολική, ηλιακή, αεροθερμική, γεωθερμική, υδροθερμική και ενέργεια των ωκεανών, υδροηλεκτρική, από βιομάζα, από τα εκλυόμενα στους χώρους υγειονομικής ταφής αέρια, από αέρια μονάδων επεξεργασίας λυμάτων και από βιοαέρια.
Krikri:
```
{
"en": "According to Directive 2009/28/EC of the European Parliament, energy from renewable non-fossil sources includes wind, solar, aerothermal, geothermal, hydrothermal and ocean energy, hydropower, biomass, landfill gas, sewage treatment plant gas and biogases.",
"es": "De conformidad con la Directiva 2009/28/CE del Parlamento Europeo, la energía procedente de fuentes renovables no fósiles incluye la eólica, solar, aerotérmica, geotérmica, hidrotérmica y la energía del océano, hidroeléctrica, procedente de la biomasa, de los gases de vertedero, de los gases de estación depuradora de aguas residuales y de los biogás.",
"fr": "Conformément à la directive 2009/28/CE du Parlement européen, les énergies renouvelables non fossiles comprennent l'éolien, le solaire, l'aérothermique, la géothermie, l'hydrothermie et l'énergie océanique, l'hydroélectricité, la biomasse, les gaz de décharge, les gaz de station d'épuration et les biogaz."
}
```

Ex. 4: Summarization: Basketball game

User: Μπορείς να δημιουργήσεις μια περίληψη 5 προτάσεων για τον παρακάτω αγώνα μπάσκετ: (User provides the text of an article describing a 2025 basketball game)
Krikri: Ο Ολυμπιακός νίκησε την Μπαρτσελόνα με 90-88 σε ένα συναρπαστικό παιχνίδι, χάρη σε ένα νικητήριο καλάθι του Φουρνιέ στην εκπνοή. Ο Βεζένκοφ και ο Γκος σημείωσαν από 24 και 20 πόντους αντίστοιχα για τους νικητές, ενώ ο Μετού ήταν ο κορυφαίος των Καταλανών με 16 πόντους. Η ομάδα του Μπαρτζώκα παρέμεινε στην κορυφή της Ευρωλίγκας μετά από αυτή τη σημαντική νίκη.

Ex. 5: History: Population exchange between Greece and Turkey

User: Συνόψισε σε μία παράγραφο την Ελληνοτουρκική ανταλλαγή πληθυσμών του 1923
Krikri: Η Ελληνοτουρκική ανταλλαγή πληθυσμών του 1923, που προβλέφθηκε από τη Συνθήκη της Λωζάνης, αποτέλεσε μια μαζική μετακίνηση πληθυσμών μεταξύ Ελλάδας και Τουρκίας με στόχο την ομογενοποίηση των εθνών-κρατών. Περίπου 1,5 εκατομμύριο Έλληνες χριστιανοί, κυρίως από την Ανατολική Θράκη, την Ιωνία, τον Πόντο και την Καππαδοκία, μετακινήθηκαν στην Ελλάδα, ενώ αντίστοιχα 500.000 Μουσουλμάνοι (κυρίως Τούρκοι αλλά και Πομάκοι και Αθίγγανοι) μετακινήθηκαν από την Ελλάδα στην Τουρκία. Αυτή η ανταλλαγή, παρά τις τραγικές της συνέπειες για τους ανθρώπους που εκτοπίστηκαν, αποτέλεσε ένα από τα σημαντικότερα γεγονότα που διαμόρφωσαν τη σύγχρονη δημογραφική και εθνοτική σύνθεση των δύο χωρών, οδηγώντας σε μια πιο ομοιογενή εθνική ταυτότητα για κάθε κράτος. Η διαδικασία αυτή ολοκληρώθηκε το 1924 και θεωρείται ορόσημο στις ελληνοτουρκικές σχέσεις.

Ex. 6: High school biology: The process of photosynthesis

User: Δώσε μια σύνοψη της διαδικασίας της φωτοσύνθεσης σε μία παράγραφο.
Krikri: Η φωτοσύνθεση είναι η διαδικασία με την οποία τα φυτά, χρησιμοποιώντας την ηλιακή ενέργεια, μετατρέπουν το διοξείδιο του άνθρακα (῝O2) και το νερό (H2O) σε γλυκόζη (῝6H12O6) και οξυγόνο (O2). Συγκεκριμένα, η χλωροφύλλη στα φύλλα απορροφά φως, διασπά το νερό απελευθερώνοντας οξυγόνο, και χρησιμοποιώντας το υδρογόνο συνθέτει γλυκόζη από το διοξείδιο του άνθρακα. Η παραγόμενη γλυκόζη αποτελεί πηγή ενέργειας για το φυτό, ενώ το οξυγόνο απελευθερώνεται στην ατμόσφαιρα ως σημαντικό παραπροϊόν, υποστηρίζοντας την επιβίωση των αερόβιων οργανισμών.

Ex. 7: Chain-of-Thought Reasoning: Math problem

System: (Appropriate system prompt used to enable thinking mode)
User: Η Αλίκη έχει ένα καλάθι με μήλα. Δίνει τα 2/5 των μήλων στην αδερφή της και το 1/5 στον αδερφό της. Τελικά της μένουν 6 μήλα. Πόσα μήλα είχε αρχικά στο καλάθι;
Krikri:

```
<thinking>
Ας λύσουμε αυτό το πρόβλημα βήμα προς βήμα:

1. Έστω x ο αρχικός αριθμός μήλων που είχε η Αλίκη
2. Γνωρίζουμε ότι:
   * Δίνει 2/5 του x στην αδερφή της
   * Δίνει 1/5 του x στον αδερφό της
   * Μετά από όλα αυτά της μένουν 6 μήλα

3. Ας κάνουμε την εξίσωση:
   * Αρχικά είχε: x μήλα
   * Της μένουν: 6 μήλα
   * Άρα: x - (2/5)x - (1/5)x = 6
   * Απλοποιώντας: x - (3/5)x = 6
   * Συνδυάζοντας όρους: (2/5)x = 6
   * Πολλαπλασιάζοντας και τα δύο μέλη με 5/2:
   * x = 15

Ας επαληθεύσουμε:
* Αρχικά είχε 15 μήλα
* Δίνει: 2/5 * 15 = 6 μήλα στην αδερφή της
* Δίνει: 1/5 * 15 = 3 μήλα στον αδερφό της
* Μένουν: 15 - 6 - 3 = 6 μήλα
Η λύση είναι σωστή!
</thinking>

<output>
Η Αλίκη είχε αρχικά 15 μήλα στο καλάθι της.

Μπορούμε να το επαληθεύσουμε:
* Έδωσε 6 μήλα στην αδερφή της (2/5 του συνόλου)
* Έδωσε 3 μήλα στον αδερφό της (1/5 του συνόλου)
* Έτσι της έμειναν 6 μήλα (15 - 6 - 3 = 6)
</output>
```