AMALIA:
A Fully Open Large Language Model for European Portuguese

Afonso Simplício[1,2], Gonçalo Vinagre[1,2], Miguel Moura Ramos[3,4], Diogo Tavares[1,2]
Rafael Ferreira[1,2], Giuseppe Attanasio[3], Duarte M. Alves[3,4], Inês Calvo[1], Inês Vieira[1]
Rui Guerra[1], James Furtado[1,2], Beatriz Canaverde[3,4], Iago Paulo[1,2], Vasco Ramos[1,2]
Diogo Glória-Silva[1,2], Miguel Faria[3], Marcos Treviso[3,4], Daniel Gomes[5], Pedro Gomes[5]
David Semedo[1,2], André Martins[3,4], João Magalhães[1,2]

[1]NOVA School of Science and Technology, [2]NOVA LINCS, [3]Instituto de Telecomunicações,
[4]Instituto Superior Técnico, Universidade de Lisboa, [5]Fundação para a Ciência e Tecnologia

Correspondence: {am.simplicio, gv.martins}@campus.fct.unl.pt

Abstract

Despite rapid progress in open large language models (LLMs), European Portuguese (pt-PT) remains underrepresented in both training data and native evaluation, with machine-translated benchmarks likely missing the variant's linguistic and cultural nuances. We introduce AMALIA, a fully open LLM that prioritizes pt-PT by using more high-quality pt-PT data during both the mid- and post-training stages. To evaluate pt-PT more faithfully, we release a suite of pt-PT benchmarks that includes translated standard tasks and four new datasets targeting pt-PT generation, linguistic competence, and pt-PT/pt-BR bias. Experiments show that AMALIA matches strong baselines on translated benchmarks while substantially improving performance on pt-PT-specific evaluations, supporting the case for targeted training and native benchmarking for European Portuguese.[1]

1 Introduction

The rapid development of large language models (LLMs) has fundamentally transformed the field of natural language processing (NLP), with open-weight models (Dubey et al., 2024; Yang et al., 2025; Kamath et al., 2025) setting new standards in tasks ranging from problem-solving to conversational AI. However, a significant limitation of current state-of-the-art models is their overwhelming focus on English-language data (Olmo et al., 2025). This imbalance is reflected in evaluation, where the scarcity of native benchmarks leads many researchers to rely on machine translation (e.g., Son et al., 2025; Xuan et al., 2025). In turn, such MT benchmarks often fail to capture the linguistic and cultural nuances of content naturally produced by native speakers (Plaza et al., 2024; Wu et al., 2025), spurring focused efforts to prioritize native, culturally specific sourcing (e.g., Attanasio et al., 2024; Singh et al., 2025). As a result, many European languages, including European Portuguese, are underrepresented in global LLMs, limiting these technologies' ability to capture the full breadth of Europe's linguistic and cultural diversity.

This paper introduces AMALIA, an LLM designed to address this imbalance by prioritizing European Portuguese and its cultural context during pretraining and post-training. Furthermore, we introduce a set of benchmarks that assess the linguistic performance and cultural sensitivity of LLMs, explicitly tailored for European Portuguese. The process of creating AMALIA began with the collection and processing of large-scale European Portuguese data and subsequent data quality filtering. We leveraged Arquivo.pt, the Portuguese Web Archive,[2] as a primary source for this data. The model was then pretrained on this data and post-trained on specific datasets for instruction following, conversational reasoning, and problem solving.

Our experiments reveal that the model is on par with other similarly-sized open models in most machine-translated benchmarks and is superior on the European Portuguese benchmarks. This is a strong indication that LLMs can indeed capture specific traits of underrepresented language varieties, even when the number of examples is orders of magnitude smaller than the dominant language variety and the dominant language, *i.e.*, Brazilian Portuguese and English, respectively.

2 Related Work

The development of fully open large language models has been dominated by English-centric training (Olmo et al., 2025), although recent efforts have been made in multilingual models (Gonzalez-Agirre et al., 2025; Martins et al., 2025; Apertus,

[1] https://github.com/AMALIA-LLM/AMALIA

[2] https://arquivo.pt/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 380–391

2025). Despite this, lower resource languages are still under-represented in these models, and language varieties distinction, such as European and Brazilian Portuguese, is still a major challenge, leading to degraded performance on language-specific tasks and cultural contexts.

To mitigate this phenomena, several efforts have addressed the lack of data and models for European Portuguese specifically. Gloria (Lopes et al., 2024) introduced an European Portuguese decoder-based LLM pretrained on a large PT-PT text corpora, and Gervasio (Santos et al., 2024) fine-tuned Llama-Instruct 2 (Dubey et al., 2024) for pt-PT.

3 Data Collection and Processing

3.1 Arquivo Data

Collections. We collected publicly available data from Arquivo.pt. Following GlórIA's (Lopes et al., 2024) efforts, we gathered WARC archives comprising a selection of collections that primarily contain general web and free books, thereby improving the variety and quantity of the Portuguese data in the pretraining phase. We obtained this data by copying directly from Arquivo.pt's data center; in its raw format, it totals 195 terabytes of WARC archives, which were subsequently processed.

Filtering. We created a data processing and filtering pipeline inspired by the FineWeb2 (Penedo et al., 2025) pipelines, using the Datatrove (Penedo et al., 2024) library. For each data collection, we begin by using URL filtering to remove any '.br' domains to reduce Brazilian Portuguese content. We also use a blacklist to remove sensitive (NSFW) content. Then, we scrape the text from the HTML files using the Trafilatura (Barbaresi, 2021) extractor and perform additional post-scraping by removing any short or duplicate lines in the text. We then followed FineWeb heuristic filters with its parameters calibrated for European Portuguese, namely the Language Identification (Kargaran et al., 2023), FineWeb Quality, and Gopher Repetition and Quality filters (Rae et al., 2022). After removing personal data and fixing remaining encoding issues, we deduplicated the data using MinHash (Broder, 1997), and labeled it according to the EuroFilter quality classification model (Martins et al., 2025). Finally, we merged the data in all collections, creating three quality splits (high, medium and low), and we used the high and medium quality, totaling 5.8 billion tokens.

GDPR compliance. The data from Arquivo.pt was collected from the public Web and respects an embargo of one year. As mentioned, we removed personal data, namely public IPs, email addresses, and phone numbers.

3.2 Long Context

Following Ramos et al. (2026), we reused the long documents from EuroLLM's pretraining corpus (Martins et al., 2025), which had previously been truncated to 4K tokens. We also collected long-context code samples from Stack-v2 (Lozhkov et al., 2024), considering only repositories with at least 500 stars and 100 forks. Additionally, we included synthetic data focused on improving recall and retrieval in long-context scenarios. Concretely, starting from publicly available datasets,[3] we filtered out samples longer than 32K tokens and removed duplicates.

3.3 Post-train Data

To improve AMALIA's instruction following capabilities, we built a data mixture targeting four categories: instruction following, conversational reasoning, mathematical problem-solving, and safety. This mixture comprises our own synthetic and manually curated data in both European Portuguese and English alongside publicly available datasets from Hugging Face. We only used data sources that had an open-source license associated to it. The complete data mixture is detailed in Table 1.

Our synthetic generation primarily followed the PersonaHub (Ge et al., 2025) approach to increase diversity in instruction following and conversational scenarios, using the PersonaHub and Nemotron personas (Meyer and Corneil, 2025) to create datasets for general, instruction following, and mathematical tasks. European Portuguese data was produced via machine translation using our pt-PT MT model (Simplício et al., 2026) and Gemma 3-27B (Kamath et al., 2025), question generation with few-shot prompting, and answer generation with Gemma 3-27B. This choice stems from our observation that Gemma 3-27B was the best available open model for European Portuguese, a finding later corroborated by experimental results (Section 6.4). We consistently relied on larger and varied models to produce prompts, questions, and

[3] Datasets available at https://huggingface.co/datasets/lvwerra/needle-llama3-16x8k and https://huggingface.co/datasets/nanotron/needle_32k_finetuning_dataset.

answers in order to reduce the risk of biased data sampling (Simplício et al., 2026).

To ensure high data quality, we applied dataset-specific filtering and formatting strategies. We removed all reasoning traces, eliminated samples containing self-referential content from other models, and utilized the DEITA Quality Scorer (Liu et al., 2023), which gives a score between 1 and 6 to filter out synthetic entries with quality scores below 5. Finally, we applied global deduplication, retaining only one entry per unique user prompt.

Instruction Following. The instruction following mixture consists of datasets created primarily via the aforementioned PersonaHub method. To address pt-PT specific nuances, we included a Portuguese Linguistic Instructions dataset, manually curated by a Portuguese linguistics expert. This dataset, comprising 200 entries, covers aspects, such as phonetics, orthography, wordplay, idiomatic expressions, and grammar classification, specific to European Portuguese.

Conversational Reasoning. This mixture provides general knowledge in a conversational format, incorporating the Persona-PT General and Persona Nemotron General datasets, and generated conversational data from Wikipedia. We also included the AMALIA Hardcoded dataset, providing 156 entries with self-referential knowledge about AMALIA's development and capabilities.

To improve system-prompt adherence (Lee et al., 2024; Mu et al., 2024), we manually selected a subset of the Hermes3 SFT dataset (Teknium et al., 2024), retaining only entries containing custom system prompts that substantially modify the model's behavior. Furthermore, from the smoltalk's smol-magpie-ultra subset (Allal et al., 2025) we kept the entries pre-labeled as excellent quality and translated the smol-summarize subset. We also processed the smoltalk2's everyday-conversations subset (Bakouch et al., 2025) by translating initial turns, generating remaining turns in European Portuguese and removing all greeting turns. We randomly selected 1M entries from the STEM split of Nemotron-Post-Training-v1 (Nathawani et al., 2025b; Bercovich et al., 2025), and removed entries with noisy artifacts. From the Nemotron-Post-Training-v2 (Nathawani et al., 2025a; NVIDIA, 2025) chat split, following RIP (Yu et al., 2025), we removed problematic instructions from the WildChat (Zhao et al., 2024) subset.

For translation capabilities, we used the PTradutor dataset (Sousa et al., 2025) for PT-EN and EN-PT translations, by selecting bidirectional translation pairs, and the WMT24++ dataset (Deutsch et al., 2025), where we crossed the XX-EN pairs with the respective PT-EN pairs to obtain XX-PT pairs and downsampled them to balance the source-target variety.

Mathematics. To increase the presence of mathematical data in European Portuguese, we generated the Persona-PT Math and Persona Nemotron Math datasets. To improve robustness to varying output formats, we modified the math split of Nemotron-Post-Training-v2 by removing the mandatory `\boxed{x}` termination from 50% of the samples, improving robustness to alternative valid output formats.

Safety. For the safety mixture, we classified every prompt of the Euroblocks SFT dataset[4] using Qwen3Guard (Zhao et al., 2025), retaining only samples that contained unsafe or controversial prompts paired with the dataset's safe responses. Additionally, we created an English safety dataset using DeepSeek-V3.2-Exp (DeepSeek-AI, 2025) that covers hierarchically organized sensitive topics and adversarial strategies (Simplício et al., 2026).

4 Model Development

To build the base AMALIA model, we modified the final pretraining phase of EuroLLM-9B (Martins et al., 2025) to improve its language modeling capabilities for pt-PT, and carried a supervised fine-tuning phase for instruction following and conversational reasoning.

4.1 Pretraining

The data sources used in AMALIA's pretraining consist of four sources. First, the original EuroLLM training mixture, which contains 40B tokens from multiple European languages—including a small portion of mathematics and code—and is largely composed of short documents rather than long sequences. To better support longer contexts, we expanded it with 60B tokens with improved coverage of the code domain, and further included 1.4B synthetically generated long-context tokens. This strategy balances shorter and longer sequences, which improves performance

[4] https://huggingface.co/datasets/utter-project/EuroBlocks-SFT-Synthetic-1124

Task	Dataset	Language	Proportion	#Tokens	License / Source
Instruction Following	Persona-PT Instruction Following	PT	0.332%	10 741 619	Apache 2.0 / Synth. Gen.
	Persona-EN Instruction Following	EN	0.807%	26 065 053	Apache 2.0 / Synth. Gen.
	Persona Nemotron Instruction Following	PT	0.147%	4 765 186	Apache 2.0 / Synth. Gen.
	Portuguese Linguistic Instructions	PT	0.001%	18 884	Apache 2.0 / Manual
Conversational Reasoning	AMALIA Hardcoded	PT & EN	0.001%	41 756	Apache 2.0 / Manual
	Persona-PT General	PT	5.945%	192 102 099	Apache 2.0 / Synth. Gen.
	Persona Nemotron General	PT	2.906%	93 916 092	Apache 2.0 / Synth. Gen.
	Wikipedia Conversations	PT	2.336%	75 493 190	Apache 2.0 / Synth. Gen.
	Nemotron-Post-Training-v1 - Chat Split	Multi	13.206%	426 741 052	CC BY 4.0 / HF
	Nemotron-Post-Training-v1 - STEM Split	EN	11.616%	375 380 057	CC BY 4.0 / HF
	Nemotron-Post-Training-v2 - Chat Split	Multi	10.993%	355 252 481	CC BY 4.0 / HF
	Nemotron-Post-Training-v2 - STEM Split	Multi	4.757%	153 720 957	CC BY 4.0 / HF
	smoltalk - Smol-Magpie-Ultra (Excellent Quality)	EN	0.118%	3 809 670	Apache 2.0 / HF
	smoltalk - Smol-summarize (Translated)	PT	1.703%	55 028 254	Apache 2.0 / HF
	smoltalk2 - everyday-conversations (Translated)	PT	0.023%	741 216	Apache 2.0 / HF
	smoltalk2 - Table-GPT	EN	0.317%	10 257 869	MIT / HF
	Hermes3 Custom ST Split	EN	2.454%	79 297 721	Apache 2.0 / HF
	PTradutor	PT $\leftrightarrow$ EN	0.740%	23 912 149	MIT / HF
	WMT'24 Multilingual Translations	Multi $\rightarrow$ PT	0.055%	1 766 630	Apache 2.0 / HF
	tulu-3-sft-olmo-2-mixture	Multi	18.540%	599 099 985	ODC-BY-1.0. / HF
Mathematics	Persona-PT Math	PT	1.256%	40 596 577	Apache 2.0 / Synth. Gen.
	Persona Nemotron Math	PT	2.847%	91 995 370	Apache 2.0 / Synth. Gen.
	Nemotron-Post-Training-v1 - Math Split	EN	12.340%	398 746 982	CC BY 4.0 / HF
	Nemotron-Post-Training-v2 - Math Split	EN	2.431%	78 563 288	CC BY 4.0 / HF
	Nemotron-Post-Training-v2 - Code Split	EN	1.656%	53 498 843	CC BY 4.0 / HF
	orca-math-word-problems-200k	EN	1.793%	57 947 703	MIT / HF
Safety	EuroBlocks Safety Samples	Multi	0.422%	13 650 597	Apache 2.0 / HF
	AI Generated Safety Data	EN	0.258%	8 346 310	Apache 2.0 / Synth. Gen.

Table 1: Data mixture used in the Supervised Fine-Tuning stage. Datasets were generated synthetically (Synth. Gen.), created manually (Manual) or pre-existent in HuggingFace (HF).

on long-context tasks while maintaining model quality on short-context tasks (Xiong et al., 2024). Additionally, to further improve European Portuguese capabilities, we included 5.8B tokens from the Arquivo.pt dataset (§3.1). We largely follow EuroLLM-9B hyperparameters, with two changes: we extend the max sequence length from 4K to 32K tokens and apply RoPE scaling (Xiong et al., 2024), increasing θ from 10,000 to 1,000,000. Training took 80 hours on 256 NVIDIA H100 GPUs.

4.2 Supervised Fine-Tuning

We fine-tuned AMALIA to improve its instruction following and conversational capabilities using the data mixture described in Table 1. Training was conducted over 14K steps (approximately 4.25 epochs), using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 10^{-5}, a cosine learning rate scheduler with a warmup ratio of 0.03, weight decay of 0.01, and `bfloat16` mixed precision. The final checkpoint was selected based on the performance on the validation set. We trained for 76 hours on 64 NVIDIA H100 GPUs.

4.3 Preference Training

For preference training, we used Direct Preference Optimization (DPO) (Rafailov et al., 2024), starting by following an approach inspired by OLMo3 (Olmo et al., 2025), using 200K prompts sampled from the SFT dataset. We experimented with generating responses using several models; the best results were obtained by sampling 32 candidate responses from our own SFT model for each prompt, scoring them with ArmoRM (Wang et al., 2024), and selecting the highest- and lowest-reward responses as the chosen and rejected answers, respectively.

While this strategy led to overall performance improvements, it resulted in declines in mathematical reasoning and instruction following. We hypothesize that these declines stem from the reward model favoring responses that do not strictly adhere to task constraints. To mitigate these issues, we adopted domain-specific data construction strategies. For our persona datasets, we used the original answers as the chosen responses and AMALIA-SFT generated answers as the rejected ones. For mathematical reasoning, we found that generating chosen responses with Qwen 3-32B (Yang

et al., 2025) and rejected responses with Qwen 3-0.6B (Yang et al., 2025) yielded the best performance. We also observed a tendency for a decline in some Portuguese-specific capabilities; therefore, we decided to add more Portuguese data to the mix, adding subsets of the Persona-Nemotron datasets, using 100K, 50K, and 30K entries from the General, Math, and Instruction Following sets, respectively, with the rejected responses generated by Qwen 3-0.6B (Yang et al., 2025).

To further improve general capabilities and explicitly target weaknesses in math and instruction following, we incorporated additional preference datasets: UltraFeedback (Cui et al., 2023), OpenAssistant 2 (Köpf et al., 2023), Abbey4799/Complex-Instructions-DPO,[5] and kira/math-dpo.[6] To improve model safety, we also included Egida-DPO-Meta-LLaMa-3.1-70B-Instruct (Garcia-Gasulla et al., 2025), HarmfulQA (Bhardwaj and Poria, 2023), and a safety dataset tailored to the Portuguese cultural context (Simplício et al., 2026). In total, this mixture comprised 478K preference pairs.

We trained for one epoch with a batch size of 128, a learning rate of 10^{-6} under a linear scheduler, and $\beta = 0.1$, taking 12 hours on 64 NVIDIA H100 GPUs.

5 Experimental Setup

Following other fully open LLMs (OLMo et al., 2024; Allal et al., 2025), we evaluate instruct and preference AMALIA variants on a suite of diverse and challenging benchmarks.

5.1 Model Baselines

We compare AMALIA to a range of instruction-tuned models of comparable size, including both open-weight and fully open-source models. Among the open-weight models, we consider Mistral-7B (Jiang et al., 2023); Ministral-8B (Jiang et al., 2024); Llama 3.1-8B (Dubey et al., 2024); Gemma 2-9B (Rivière et al., 2024); Gemma 3-12B (Kamath et al., 2025), all designed for broad cross-lingual generalization; and the Qwen family, comprising Qwen 2.5-7B (Yang et al., 2024) and Qwen 3-8B (Yang et al., 2025), which are known for strong multilingual and reasoning capabilities. For fully open-source models that release both code and data, we evaluate OLMo 2-7B (OLMo et al., 2024), a strong English-only baseline; Salamandra-7B (Gonzalez-Agirre et al., 2025), focused on Iberian languages; EuroLLM-9B (Martins et al., 2025), tailored to European languages; and Apertus-8B (Apertus, 2025), trained on hundreds of languages. Additionally, we include Gervasio-8B,[7], a Llama-Instruct 3.1-8B model, fine-tuned for pt-PT.

5.2 PT Benchmark Collection

Benchmarks were selected according to the following criteria: (1) originally authored in European Portuguese (pt-PT) or Brazilian Portuguese (pt-BR), (2) translated by humans, and (3) existing machine-translated datasets. To further expand coverage of evaluation capabilities, we additionally machine-translated benchmarks lacking a Portuguese variant and compiled new corpora to assess pt-PT relevant skills (Simplício et al., 2026). In all evaluations, every prompt component, including questions, answers, and instructions, were presented in pt-PT.

5.3 pt-PT Benchmark Collection

Developing datasets in pt-PT requires special attention to cultural and linguistic distinctions from pt-BR. To address these differences, we introduce four new datasets designed to support benchmarking and evaluation for pt-PT language models.

PT-PT Completions (PT-C). This task evaluates a model's ability to complete sentences in European Portuguese (pt-PT). Each sentence has a blank and two options: one in pt-PT and one in pt-BR, where models should select the first option by default. The dataset contains 70 manually curated examples highlighting pt-PT/pt-BR differences, e.g., *"Vou à estação de ____ para comprar um bilhete."*, with options (A) "comboios"—a pt-PT term—and (B) "trem"—a pt-BR term.

PT Exams (PT-E) (Tavares et al., 2026). We extract questions and answers from the official Portuguese national high school exams.[8] The dataset comprises 1.8K multiple-choice and 1.8K open questions from 2006–2023, over six subjects: Mathematics, Portuguese, History, Geography, Biology/Geology, and Philosophy. All questions were

[5] https://huggingface.co/datasets/Abbey4799/Complex-Instructions-DPO

[6] https://huggingface.co/datasets/kira/math-dpo

[7] https://huggingface.co/PORTULAN/gervasio-8b-portuguese-ptpt-decoder

[8] https://github.com/AMALIA-LLM/pheb

Dataset	Category	Source Language	CoT	#Shots	Metric	Reference
ALBA	PT Grammar	🇵🇹		0	LLM-Judge	New
PT Completions	pt-PT Generation	🇵🇹		0	Accuracy	New
PT Exams	General Knowledge	🇵🇹	✓	0	Accuracy	New
PT Exams Open Questions	General Knowledge	🇵🇹	✓	0	LLM-Judge	New
P3B3	pt-PT/BR Generation	🇵🇹		0	pt-PT Level	New
FRMT	Translation	🇵🇹		5	chrF	(Riley et al., 2023)
ARC-C	Commonsense Reasoning	🇬🇧		0	Accuracy	(Thellmann et al., 2024)
GSM8K	Mathematics	🇬🇧	✓	8	Exact Match	(Thellmann et al., 2024)
MMLU	General Knowledge	🇬🇧		0	Accuracy	(Thellmann et al., 2024)
TruthfulQA	NLU	🇬🇧		6	Accuracy	(Thellmann et al., 2024)
PIQA	Commonsense Reasoning	🇬🇧		0	Accuracy	New (MT)
SIQA	Commonsense Reasoning	🇬🇧		0	Accuracy	New (MT)
IFEval	Instruction Following	🇬🇧		0	Prompt level strict	New (MT)
BBH	Reasoning	🇬🇧	✓	3	Exact Match	New (MT)
Simple Safety Tests	Safety	🇬🇧		0	ASR	New (MT)
XSTest	Safety	🇬🇧		0	ASR	New (MT)
Multilingual ADV bench	Safety	🇬🇧		0	ASR	simonycl/multilingual_advbench

Table 2: Summary of datasets used for evaluation. **New** denotes novel datasets created for pt-PT evaluation. **New (MT)** are datasets newly translated into pt-PT via our pt-PT MT model (Simplício et al., 2026). Note: Some BBH tasks were removed from pt-PT evaluation as they do not translate correctly to Portuguese (e.g. hyperbaton task).

written and curated by educators to ensure correctness and alignment with the national curriculum. These questions assess factual knowledge, reasoning skills, and language comprehension.

ALBA (Vieira et al., 2026). Automated Linguistics Benchmark for baseline Assessment (ALBA) is a pt-PT linguistics benchmark designed to address gaps in existing language evaluation resources. Developed manually by domain experts from our team, it covers eight categories: language variety, culture-bound semantics, discourse analysis, wordplay, syntax, morphology, lexicology, and phonetics & phonology. The dataset comprises 800 questions (100 per category), of which 30 are paired with three graded reference answers to enable validation against human judgment.

P3B3.[9] There are many nuances to consider when comparing pt-PT to pt-BR, particularly with respect to grammatical structures and lexical variation. To evaluate potential model bias toward either variant, we construct the pt-PT–pt-BR Bias Benchmark (P3B3), comprising 200 multi-turn conversational prompts, each with 2 to 5 turns. The prompts are variant-agnostic and designed to provoke a variant-specific response. At each turn, the model generates a response using the accumulated dialogue history. We then use automated methods to assess whether responses align more with pt-PT or pt-BR.

[9] https://github.com/AMALIA-LLM/p3b3-benchmark

5.4 Benchmark Selection

We evaluate AMALIA on a diverse set of tasks following OLMo et al. (2024); Allal et al. (2025). Specifically, we include ARC Challenge (ARC-C; Clark et al., 2018), MMLU (Hendrycks et al., 2021), TruthfulQA (TQA; Lin et al., 2022), PiQA (Bisk et al., 2020), and SiQA (Sap et al., 2019). Together, these tasks assess a broad spectrum of capabilities, including commonsense reasoning, scientific and world knowledge, reading comprehension, question answering, and cloze-style completion. Additionally, we test models on more complex instruction following datasets, such as BigBenchHard (BBH; Srivastava et al., 2023), IFEval (IFE; Zhou et al., 2023), along with GSM8k (Cobbe et al., 2021) for math reasoning.

To more directly evaluate Portuguese-specific performance, we extend this suite with benchmarks designed to test linguistic and domain knowledge in Portuguese. These include the pt-PT translation dataset FRMT (Riley et al., 2023), and our newly created pt-PT–specific datasets from Section 5.3.

We also include safety-focused benchmarks. Specifically, we evaluate on Simple Safety Tests (SST; Vidgen et al., 2023), a corpus of unsafe prompts spanning five harm categories; the unsafe subset of XSTest (XST; Röttger et al., 2024), and the Multilingual ADV Bench (MAD; Zou et al., 2023), which contains unsafe prompts covering a broad range of adversarial attack types.

Model	ARC-C	MMLU	TQA	GSM8K	IFEval	PT-E (CoT)	PT-C	PiQA	SiQA	FRMT	BBH
Fully open models											
OLMo 2-7B	47.8	42.6	50.8	51.6	50.6	45.0	32.9	61.0	41.9	61.0	37.0
Salamandra-7B	52.2	47.1	52.6	18.7	24.2	37.7	44.3	70.4	44.8	65.6	36.3
EuroLLM-9B	71.2	52.7	54.9	52.5	49.4	55.2	58.6	71.7	40.8	70.4	47.6
Apertus-8B	68.3	57.1	57.9	48.8	58.6	55.7	40.0	**73.4**	43.2	67.2	48.3
AMALIA-9B-SFT	77.9	55.4	57.0	58.8	56.7	63.1	**71.4**	71.7	44.7	**70.5**	50.3
AMALIA-9B-DPO	78.9	58.8	**63.5**	52.4	61.6	68.4	67.1	72.5	**46.3**	70.2	47.3
Open weight models											
Llama 3.1-8B	75.8	58.5	58.5	68.0	55.1	66.7	38.6	68.5	43.4	66.3	68.5
Gervasio-8B	79.0	59.8	56.8	67.3	55.6	66.0	30.0	69.1	44.8	66.3	68.1
Gemma 2-9B	85.6	65.8	62.7	73.1	47.7	70.9	34.3	70.3	41.0	66.5	50.8
Gemma 3-12B	86.9	**68.0**	63.2	**79.2**	71.7	81.3	21.4	69.0	43.8	69.5	**79.3**
Qwen 2.5-7B	83.7	65.7	63.2	74.0	61.6	73.7	40.0	66.6	40.7	63.9	56.5
Qwen 3-8B	**87.4**	64.9	57.2	77.6	**72.1**	**82.8**	28.6	64.7	41.2	65.1	23.3
Mistral-7B	61.8	51.0	61.1	38.3	39.7	53.6	34.3	66.9	40.9	63.3	52.1
Ministral-8B	73.6	55.5	56.0	66.9	44.0	64.9	40.0	70.7	42.3	66.1	59.1

Table 3: Results of instruction-tuned variants. Bold represents best model and underline best fully open model.

5.5 Benchmark Configuration

We performed evaluations using a combination of LM-Eval-Harness,[10] an evaluation framework that supports custom task development, and custom scripts for LLM-as-a-Judge evaluation. Inference was conducted on a single NVIDIA H100 GPU, with models loaded in `bf16` precision and served through vLLM (Kwon et al., 2023), using automatic batch sizing and a maximum context length of 4096 tokens. Whenever available, we used the official task generation parameters provided by each benchmark and used greedy decoding when such parameters were not specified. Following OLMES standards (Gu et al., 2025), tasks are formatted either as multiple-choice or chain-of-thought prompting. We evaluate models in zero- and few-shot settings and report the metric associated with each dataset.

Table 2 summarizes the full benchmark configuration. For each dataset, we include its source and translation method. In addition to the newly constructed datasets tailored for pt-PT evaluation, several datasets were automatically translated using a pt-PT translation model (Simplício et al., 2026), which is explicitly optimized for pt-PT translation, since models tend to translate to pt-BR.

6 Results and Discussion

6.1 Instruct Model Evaluation

We show the performance of instruction-tuned variants of AMALIA on multiple tasks in Table 3. Overall, AMALIA achieves state-of-the-art performance among fully open models of comparable size across most benchmarks. The SFT variant leads on four benchmarks, exhibiting particularly strong performance in Portuguese language understanding and grade school mathematics. Building on this foundation, the preference-tuned (DPO) version outperforms all other fully open models on six tasks, with especially notable improvements on PT-E, reflecting a strong grasp of the Portuguese school curriculum. In addition, AMALIA's performance on ARC-C, MMLU, TQA, PiQA, SiQA indicate that it acquired good general and commonsense knowledge, and IFEval results demonstrate the models flexibility across varied instruction-following scenarios.

6.2 PT-Exams: Open Questions

After evaluating the model's knowledge using the MCQ from PT-Exams (Section 5.3), we further assess its ability to comprehend and produce written pt-PT. To this end, we again use the national Portuguese exams, but now focus on the open-ended questions for the Portuguese subject, totaling 336 questions. Unlike MCQs, these questions do not have a single correct answer. Instead, the official exams provide correction rubrics that specify both the expected content and the quality of the written response. We extract the exam questions and evaluation criteria for the Portuguese subject and use Gemini 2.5 Pro to assess model-generated answers according to these rubrics. This evaluator was selected due to its strong agreement with human judgments (Tavares et al., 2026).

Table 4 shows that AMALIA achieves the high-

[10] https://github.com/EleutherAI/lm-evaluation-harness

Model	PT-Exams Open Questions	ALBA	P3B3
Fully open models			
OLMo 2-7B	43.0	16.9	18.6
Salamandra-7B	34.1	27.4	42.7
EuroLLM-9B	56.1	38.5	70.5
Apertus-8B	54.7	38.7	28.1
AMALIA-9B-SFT	62.0	37.1	91.3
AMALIA-9B-DPO	66.0	43.6	**95.9**
Open weight models			
Llama 3.1-8B	53.8	31.3	27.8
Gervasio-8B	53.2	31.1	24.7
Qwen 2.5-7B	56.8	31.0	20.0
Qwen 3-8B	**77.3**	49.8	18.9
Gemma 2-9B	69.7	41.1	72.1
Gemma 3-12B	76.6	**51.1**	88.3
Mistral-7B	44.6	21.7	19.2
Ministral-8B	62.0	35.6	22.1

Table 4: Language generation results of instruction-tuned variants on the newly created pt-PT benchmarks.

Model	ADV-Bench ↓	SST ↓	XSTest ↓
Fully open models			
OLMo 2-7B	6.9	20.4	17.4
Salamandra-7B	6.6	18.4	23.4
EuroLLM-9B	1.4	3.4	1.9
Apertus-8B	2.7	4.4	3.9
AMALIA-9B-SFT	1.9	3.4	1.9
AMALIA-9B-DPO	**0.8**	1.4	**0.9**
Open weight models			
Llama 3.1-8B	10.8	7.4	2.9
Gervasio-8B	12.9	9.4	3.9
Qwen 2.5-7B	1.8	3.4	2.4
Qwen 3-8B	2.1	4.4	4.4
Gemma 2-9B	1.0	**0.4**	1.4
Gemma 3-12B	2.9	2.4	3.4
Mistral-7B	34.4	19.4	13.4
Ministral-8B	18.3	8.4	9.9

Table 5: ASR results of instruction-tuned variants on safety benchmarks assessed by Qwen3Guard-8B.

est performance among fully open models with significant increases with DPO, demonstrating strong Portuguese text understanding and clear, grammatically correct language generation.

6.3 ALBA: pt-PT Linguistics Evaluation

We evaluate pt-PT linguistic competence using the ALBA dataset (Section 5.3), which covers a broad range of linguistic phenomena. Evaluation is performed using Gemini 2.5 Pro as an automatic judge, producing scores on a 1–5 scale that are rescaled to a 0–100 range. Results in Table 4 show that AMALIA-DPO achieves the strongest performance among fully open models.

Drilling down on individual categories (Vieira et al., 2026), we observed that AMALIA-DPO achieves the best results in both Lexicology and Culture-bound Semantics, surpassing Gemma 3 and highlighting its above-average understanding of pt-PT–specific linguistic skills. In contrast, we found that Phonology and Wordplay emerge as the most challenging categories, reflecting the intrinsic difficulty of these more complex and nuanced tasks. Despite these variations, AMALIA-DPO showcases a solid performance across a broad range of pt-PT linguistic traits.

6.4 P3B3: Model Bias and pt-PT Generation

We assess the models' ability to generate pt-PT using the P3B3 dataset (Section 5.3) by explicitly instructing the model to respond in pt-PT. This setup allows us to evaluate both controllability and fidelity to the specified language variant. For automatic evaluation, we use Gemini 2.5-Pro with a specialized prompt that analyzes linguistic features, produces a short rationale, and assigns a score, with scores showing strong agreement with human judgments (Simplício et al., 2026).

Table 4 shows that most models exhibit a strong bias toward pt-BR, reflecting its higher-resource status. Even though models were explicitly prompted to use pt-PT, only AMALIA, EuroLLM, and Gemma are able to consistently write in pt-PT. In specific, both versions of AMALIA achieve the highest pt-PT scores on P3B3, indicating a strong ability produce outputs that conform to salient linguistic features of European Portuguese.

6.5 Safety Results

To evaluate safety, we employ a set of safety benchmarks covering both commonsense safety and jailbreak scenarios. We report the attack success rate (ASR) as assessed by Qwen3Guard-8B (Zhao et al., 2025). An attack is considered successful if the classifier labels the model's response as either "unsafe" or "controversial".

As shown in Table 5, all AMALIA variants exhibit strong safety performance, achieving consistently low ASR across benchmarks. Notably, AMALIA-DPO achieves the lowest ASR among all models on two datasets and delivers the best performance among fully open models on SST, indicating a strong ability to resist jailbreak attempts. Furthermore, our DPO step further strengthened the safety of the SFT model, cutting its ASR by more than half across the three reported benchmarks.

7 Conclusions

This paper presents AMALIA, an LLM that prioritizes the European Portuguese language and its cultural context. AMALIA leverages Arquivo.pt data and post-training data specifically curated for European Portuguese, and was trained in a three-stage process: pretraining, supervised fine-tuning, and preference tuning. To address the scarcity of suitable pt-PT benchmarks, we developed dedicated benchmarks tailored to European Portuguese, designed to capture linguistic and cultural nuances. Furthermore, we translated existing benchmarks from EN→pt-PT using a dedicated translation model, and then complement the suite by introducing a national high-school pt-PT exams benchmark. Our results show that AMALIA outperforms previously released fully-open models on European Portuguese. In language understanding and reasoning, the model achieves state-of-the-art or comparable performance, while on language generation tasks, it excels in the quality and fluency of pt-PT text. Safety evaluations further indicate that the model is aligned with state-of-the-art standards.

8 Future Work

Although AMALIA represents a significant step toward better fully-open European Portuguese language models, there are areas that require a continued development and improvement. One of the main challenges is the scarcity of pt-PT datasets for both training and evaluation, making it difficult to improve and assess capabilities on Portuguese-specific cultural knowledge and tasks. Another challenge is limited computational resources for pre-training, where compared to larger models we are limited to a few Trillion tokens.

We plan to continue developing new training data mixtures aimed at improving reasoning capabilities in pt-PT and pt-BR, to extend the context window, and to experiment with additional post-training approaches such as curriculum learning and reinforcement learning with verifiable rewards. We also aim to further enhance AMALIA's Portuguese cultural knowledge by developing targeted training data. Expanding multilingual capabilities and introducing tool-calling functionality are other priorities for future iterations. Finally, we aim to continuously expand our benchmark suite with new pt-PT and pt-BR evaluations covering core tasks such as instruction following, mathematical reasoning, and linguistic competence, alongside assessments of cultural and regional knowledge.

Acknowledgments

This work was supported by the AMALIA project under Measure RE-C05-i08 of the Portuguese national Programa de Recuperação e Resiliência. We also acknowledge the support of Fundação para a Ciência e Tecnologia (FCT) and the NOVA LINCS project (UID/04516/2025). Finally, we thank the Barcelona Supercomputing Center (BSC) for providing the computational resources that made this work possible. The IT team is supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

Loubna Ben Allal, Anton Lozhkov, and 20 others. 2025. Smollm2: When smol goes big - data-centric training of a small language model. *ArXiv*, abs/2502.02737.

Project Apertus. 2025. Apertus: Democratizing open and compliant llms for global language environments. *CoRR*, abs/2509.14233.

Giuseppe Attanasio, Pierpaolo Basile, and 9 others. 2024. CALAMITA: challenge the abilities of language models in italian. In *CLiC-it 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elie Bakouch, Loubna Ben Allal, and 21 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *ACL 2021 - System Demonstrations*, pages 122–131. ACL.

Akhiad Bercovich, Itay Levy, and 131 others. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *Preprint*, arXiv:2308.09662.

Yonatan Bisk, Rowan Zellers, and 3 others. 2020. PIQA: reasoning about physical commonsense in natural language. In *AAAI 2020*, pages 7432–7439. AAAI Press.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997*, pages 21–29.

Peter Clark, Isaac Cowhey, and 5 others. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, and 10 others. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Ganqu Cui, Lifan Yuan, and 7 others. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

Daniel Deutsch, Eleftheria Briakou, and 15 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *Preprint*, arXiv:2502.12404.

Abhimanyu Dubey, Abhinav Jauhri, and 9 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dario Garcia-Gasulla, Adrián Tormos, and 5 others. 2025. Efficient safety retrofitting against jailbreaking for llms. In *SAFECOMP 2025 Workshops*, volume 15955 of *Lecture Notes in Computer Science*, pages 537–565. Springer.

Tao Ge, Xin Chan, and 4 others. 2025. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Aitor Gonzalez-Agirre, Marc Pàmies, and 9 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Yuling Gu, Oyvind Tafjord, and 4 others. 2025. OLMES: A standard for language model evaluations. In *Findings of the ACL: NAACL 2025*, pages 5005–5033. ACL.

Dan Hendrycks, Collin Burns, and 5 others. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Albert Jiang, Alexandre Abou Chahine, and Others. 2024. Ministral 8b instruct 2410. `https://huggingface.co/mistralai/Ministral-8B-Instruct-2410`.

Albert Q. Jiang, Alexandre Sablayrolles, and 16 others. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Aishwarya Kamath, Johan Ferret, and 8 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Amir Kargaran, Ayyoob Imani, and 2 others. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 6155–6218. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, and 7 others. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP 2023*, pages 611–626. ACM.

Andreas Köpf, Yannic Kilcher, and 16 others. 2023. Openassistant conversations – democratizing large language model alignment. *Preprint*, arXiv:2304.07327.

Seongyun Lee, Sue Hyun Park, and 2 others. 2024. Aligning to thousands of preferences via system message generalization. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL 2022*, pages 3214–3252.

Wei Liu, Weihao Zeng, and 3 others. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *Preprint*, arXiv:2312.15685.

Ricardo Lopes, João Magalhães, and David Semedo. 2024. Glória: A generative and open large language model for portuguese. In *PROPOR 2024, Volume 1*, pages 441–453. ACL.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Anton Lozhkov, Raymond Li, and 55 others. 2024. Starcoder 2 and the stack v2: The next generation. *CoRR*, abs/2402.19173.

Pedro Henrique Martins, João Alves, and 15 others. 2025. Eurollm-9b: Technical report. *Preprint*, arXiv:2506.04079.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions.

Norman Mu, Jonathan Lu, and 2 others. 2024. A closer look at system message robustness. In *Neurips Safe Generative AI Workshop 2024*.

Dhruv Nathawani, Shuoyang Ding, and 6 others. 2025a. Nemotron-Post-Training-Dataset-v2.

Dhruv Nathawani, Igor Gitman, and 6 others. 2025b. Nemotron-Post-Training-Dataset-v1.

NVIDIA. 2025. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *Preprint*, arXiv:2508.14444.

Team Olmo, Allyson Ettinger, and 66 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Team OLMo, Pete Walsh, and 38 others. 2024. 2 olmo 2 furious. *ArXiv*, abs/2501.00656.

Guilherme Penedo, Hynek Kydlíček, and 3 others. 2024. Datatrove: large scale data processing.

Guilherme Penedo, Hynek Kydlíček, and 8 others. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Irene Plaza, Nina Melero, and 5 others. 2024. Spanish and llm benchmarks: Is mmlu lost in translation? *GACLM*, pages 104–108.

Jack W. Rae, Sebastian Borgeaud, and 78 others. 2022. Scaling language models: Methods, analysis & insights from training gopher. *Preprint*, arXiv:2112.11446.

Rafael Rafailov, Archit Sharma, and 4 others. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Miguel Moura Ramos, Duarte M. Alves, and 16 others. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Parker Riley, Timothy Dozat, and 6 others. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Trans. Assoc. Comput. Linguistics*, 11:671–685.

Morgane Rivière, Shreya Pathak, and 97 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Paul Röttger, Hannah Kirk, and 4 others. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *NAACL 2024 (Volume 1: Long Papers)*, pages 5377–5400. ACL.

Rodrigo Santos, João Ricardo Silva, and 3 others. 2024. Advancing generative AI for Portuguese with open decoder gervásio PT*. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 16–26, Torino, Italia. ELRA and ICCL.

Maarten Sap, Hannah Rashkin, and 3 others. 2019. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

Afonso Simplício, Gonçalo Vinagre, and 20 others. 2026. AMALIA Technical Report: A Fully Open Source Large Language Model for European Portuguese.

Shivalika Singh, Angelika Romanou, and 21 others. 2025. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In *ACL 2025*, pages 18761–18799. ACL.

Guijin Son, Hanwool Lee, and 7 others. 2025. KMMLU: measuring massive multitask language understanding in korean. In *NAACL 2025*, pages 4076–4104. ACL.

Hugo Sousa, Satya Almasian, and 2 others. 2025. Tradutor: Building a variety specific translation model. *AAAI*, 39(24):25183–25191.

Aarohi Srivastava, Abhinav Rastogi, and 448 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Diogo Tavares, Rafael Ferreira, and 7 others. 2026. PHEB: An european portuguese high school-level llm benchmark. In *Proceedings of the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026)*. ELRA and ICCL.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *Preprint*, arXiv:2408.11857.

Klaudia Thellmann, Bernhard Stadler, and 9 others. 2024. Towards multilingual LLM evaluation for european languages. *CoRR*, abs/2410.08928.

Bertie Vidgen, Hannah Rose Kirk, and 5 others. 2023. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *CoRR*, abs/2311.08370.

Inês Vieira, Inês Calvo, and 7 others. 2026. ALBA: A european portuguese benchmark for evaluating language and linguistic dimensions in generative llms. In *PROPOR 2026*.

Haoxiang Wang, Wei Xiong, and 3 others. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the ACL: EMNLP 2024*, pages 10582–10592. ACL.

Minghao Wu, Weixuan Wang, and 8 others. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *ArXiv*, abs/2504.15521.

Wenhan Xiong, Jingyu Liu, and 19 others. 2024. Effective long-context scaling of foundation models. In *NAACL 2024*, pages 4643–4663. ACL.

Weihao Xuan, Rui Yang, and 16 others. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *ArXiv*, abs/2503.10497.

An Yang, Anfeng Li, and 58 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, and 39 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Ping Yu, Weizhe Yuan, and 5 others. 2025. R.i.p.: Better models by survival of the fittest prompts. *Preprint*, arXiv:2501.18578.

Haiquan Zhao, Chenhan Yuan, and 41 others. 2025. Qwen3guard technical report. *CoRR*, abs/2510.14276.

Wenting Zhao, Xiang Ren, and 4 others. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Jeffrey Zhou, Tianjian Lu, and 6 others. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Andy Zou, Zifan Wang, and 2 others. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.