MATH-PT: A Math Reasoning Benchmark for European and Brazilian Portuguese

Tiago Teixeira[1], Ana Carolina Erthal[2], Juan Belieni[2], Beatriz Canaverde[1,3], Diego Mesquita[2], Miguel Faria[3], Eliezer de Souza da Silva[4,5], André F. T. Martins[1, 3]

[1]Instituto Superior Técnico, Universidade de Lisboa, [2]Fundação Getulio Vargas, [3]Instituto de Telecomunicações, [4]Universidade de Coimbra, CISUC/LASI, DEI [5]Basque Center for Applied Mathematics

{tiago.renou, beatriz.canaverde, miguel.faria, andre.t.martins}@tecnico.ulisboa.pt,
{acarolerthal, juanbelieni}@gmail.com, diego.mesquita@fgv.br, eliezer.silva@uc.pt

Abstract

The use of large language models (LLMs) for complex mathematical reasoning is an emergent area of research, with fast progress in methods, models, and benchmark datasets. However, most mathematical reasoning evaluations exhibit a significant linguistic bias, with the vast majority of benchmark datasets being exclusively in English or (at best) translated from English. We address this limitation by introducing MATH-PT, a novel dataset comprising 1,729 mathematical problems written in European and Brazilian Portuguese. MATH-PT is curated from a variety of high-quality native sources, including mathematical Olympiads, competitions, and exams from Portugal and Brazil. We present a comprehensive benchmark of current state-of-the-art LLMs on MATH-PT, revealing that frontier reasoning models achieve strong performance in multiple choice questions compared to open weight models, but that their performance decreases for questions with figures or open-ended questions. To facilitate future research, we release the benchmark dataset[1] and model outputs.[2]

1 Introduction

Evaluating mathematical reasoning has become a key challenge in large language model (LLM) research, motivating a growing ecosystem of benchmarks such as MATH (Hendrycks et al., 2021), MathVista (Lu et al., 2023), MathBench (Liu et al., 2024), and MathArena (Balunovic et al., 2025). However, these resources share a major limitation: they are almost exclusively written in English, and multilingual variants typically resort to translations of English benchmarks (Shi et al., 2023; Son et al., 2025; Wang et al., 2025). This creates a linguistic bias that obscures whether LLM mathematical proficiency genuinely transfers across languages.

Existing Portuguese NLP benchmarks, on the other hand, cover a range of tasks—general evaluation suites (Paula et al., 2024; Scalercio et al., 2025; Baucells et al., 2025), toxicity detection (da Silva Oliveira et al., 2024), hate speech explanation (Salles et al., 2025), truthfulness (Calvo Figueras et al., 2025), multi-factor explanations (Trager et al., 2025), and common-sense reasoning (Ponti et al., 2020). Yet, no benchmark targets mathematical reasoning, and none leverage the rich ecosystem of mathematical Olympiads and exams from Portugal and Brazil.

To address this gap, we introduce MATH-PT, the first native Portuguese benchmark for mathematical reasoning. Our contributions are threefold:

- We introduce a new evaluation dataset natively in Portuguese, encompassing both the European and Brazilian variants. We curate **1,729 multiple-choice and open-ended math questions** from high-quality, originally Portuguese sources—including Olympiads and school exams from Portugal and Brazil, covering levels from primary to pre-university.
- We benchmark **13 frontier and open-source LLMs**, with and without reasoning capabilities, revealing variations across dialects, difficulty levels, and problem types.
- We show empirically that frontier models have strong performance on this task, but that accuracy varies significantly with **model size** and **question difficulty** (level, with/without figures, multiple choice/open ended), with significant room for improvement.

2 Dataset Creation

European Portuguese The questions in European Portuguese are sourced from the Portuguese Mathematical Olympiad (*Olimpíadas Portuguesas da Matemática*, OPM), a yearly mathematics competition organized by *Sociedade Portuguesa*

[1] https://huggingface.co/datasets/tiagoteixeira03/MATH-PT

[2] https://github.com/deep-spin/math-benchmark

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1005–1010

Country	Competition	Level	MC	OE	Total
Portugal	**OPM**	1	78	34	112
		2	169	96	265
		3	143	193	336
		4	29	221	250
Brazil	**OBMEP**	2	118	–	118
		3	140	–	140
		4	167	–	167
	OBM	2	–	32	32
		3	–	58	58
		4	–	38	38
	OMIF	4	103	–	103
	ELLM	5	53	–	53
	ITA	5	57	–	57
Total			**1,057**	**672**	**1,729**

Table 1: Number of questions per competition level and type across European and Brazilian Portuguese.

de Matemática (SPM), which kindly provided the available LaTeX files covering all editions of the competition from the 1997–98 to the 2024–25 academic years. The OPM are structured into four levels, each corresponding to specific school grades: **Level 1** (*Pre-Olympiad*, 5th grade), **Level 2** (*Junior Category*, 6th–7th grades), **Level 3** (*Category A*, 8th–9th grades), and **Level 4** (*Category B*, 10th–12th grades). The Junior, A, and B categories each include two elimination rounds and a national final, whereas the *Pre-Olympiad* consists of a single round designed to encourage participation in future editions. Each exam might feature **multiple-choice questions** (**MC**), **open-ended questions** (**OE**), or both. A non-negligible subset of the questions (316 out of 963) contains **visual elements**, particularly geometry-related problems. These figures are generated directly in LaTeX using environments such as `picture`, `PStricks`, or `array`. For those questions, we preserve the original LaTeX code and include it directly in the question to maintain the integrity of the visual information. Similarly, when a question statement contains tables (encoded using the `tabular` environment), we preserve the LaTeX code, as this provides the most faithful and structured representation of tabular data for LLMs.

All mathematical expressions are kept in their original LaTeX form, delimited by `$$`, following standard LaTeX notation. All other LaTeX markup was converted to plain text using the `pylatexenc` library, ensuring that the textual content remained readable while retaining mathematical fidelity. Statistics of the dataset are provided in Table 1.

Brazilian Portuguese We collect questions from large-scale national competitions: the Brazilian Mathematical Olympiad of Public Schools (*Olimpíada Brasileira de Matemática das Escolas Públicas*, OBMEP), the Mathematics Olympiad of Federal Institutions (*Olimpíada de Matemática das Instituições Federais*, OMIF), the Elon Lages Lima Mathematics Competition (*Competição Elon Lages Lima de Matemática*, ELLM) and the Entrance Exam for the Technological Institute of Aeronautics (*Vestibular do Instituto Tecnológico de Aeronáutica*, ITA). The competitions are organized into distinct levels: **Level 2** (6th–7th grades), **Level 3** (8th–9th grades), **Level 4** (10th–12th grades) and **Level 5** (pre-university), where OBMEP covers levels 2–4, OMIF covers level 4, and both ELLM and ITA target level 5. Some competitions include both multiple-choice and open-ended exams.

In all cases, LaTeX sources are not publicly available, only PDF versions of the exams and official answer keys are distributed. Thus, we adopt an automatic extraction pipeline employing a smaller language model (`gpt-5-mini`) to read the PDF content and produce a structured representation of each exam. This approach was adopted due to the limitations of standard PDF-to-text conversion tools in accurately recovering the mathematical structures required for question solving. Concretely, each exam PDF is first converted to page-level images using standard PDF tools. Then, each page is provided to `gpt-5-mini` with prompts instructing the model to (i) segment the page into individual questions; (ii) identify, for each question, the statement, the list of answer choices, and whether any visuals were necessary to solve the question; and (iii) normalize the output into a machine-readable schema (JSON). The prompts explicitly enforce constraints such as the expected number of options (typically five, labeled A–E), the need to represent all mathematical expressions in LaTeX delimited by `$$` followed by standard LaTeX notation, and the requirement that every extracted question be self-contained. Such competitions include questions that critically depend on visual information. In the PDFs, such content may appear as embedded raster images, vector diagrams, or tables. When the figure was provided through a clearly structured table (e.g., a small numeric grid or a simple schedule), `gpt-5-mini` was instructed to represent it using a minimal `tabular` environment, which we preserved in the final text. Questions with more complex diagrams and images, where visual con-

```
Resolve a seguinte questão de escolha múltipla de
matemática. Certifica-te de colocar a letra da opção
correta (A,B,C,D ou E), e apenas a letra da opção correta,
dentro de \\boxed{}. Usa Português Europeu para pensar e
responder.
Questão: Quantos fósforos são precisos para fomar o
oitavo termo da seguinte sequência: (image1)
Conteúdo referenciado:
[image1]
\begin{picture}(...)\end{picture}
A) 21 B) 24 C) 27 D) 30 E) 34
```

Figure 1: Example of a European Portuguese (pt-PT) prompt for a multiple-choice question with a figure.

```
Resolva a seguinte questão aberta de matemática.
Certifique-se de colocar a resposta final dentro de
\\boxed{}. Use Português do Brasil para pensar e
responder.
Questão: Qual o menor valor de $n$ para que um polígono
com $n$ lados tenha a soma de seus ângulos internos maior
que $2012$ graus?
```

Figure 2: Example of a Brazilian Portuguese (pt-BR) prompt used for an open-ended question.

tent could not be reliably represented in plain text or LaTeX, were discarded from the benchmark.

3 Model Evaluation

To ensure a consistent and reproducible evaluation protocol, we adopted a standardized prompting strategy adapted to each linguistic variant (European and Brazilian Portuguese) and to the presence or absence of visual elements in the question.

For the multiple-choice questions, we use a direct instruction prompt in Portuguese (Figure 1). The prompt explicitly instructs the model to: (i) solve the multiple-choice mathematics question; (ii) output only the letter of the correct option (A, B, C, D, or E); and (iii) place that letter inside a `\boxed{}` command. If the question contains figures (European Portuguese only), we include an additional block enumerating the referenced visual content, which contains the LaTeX source code with environments such as `picture`, `PStricks`, or `array` extracted from the source exam. For the Brazilian Portuguese subset (pt-BR), which consists exclusively of questions without figures, we adopt an analogous prompt written in Brazilian Portuguese. For open-ended questions, instead of asking for the letter of the correct option, we simply ask the model to make sure to place the final answer inside a `\boxed {}` command (Figure 2).

All models are evaluated in a zero-shot setting, without chain-of-thought supervision or few-shot examples. Each question is submitted independently to the model using the appropriate

```
You are an impartial judge. Your task is to evaluate the
correctness of a model's answer to an open-ended math
question by comparing it ONLY to the gold solution.
The model's answer may use different reasoning steps or
formatting than the gold solution; this is acceptable
**as long as the final mathematical result is
equivalent**. Focus strictly on mathematical correctness.
Instructions:
1. Compare the **final result** of the model answer to
the gold answer.
2. Minor algebraic, arithmetic, or formatting
differences are allowed.
3. Completely ignore writing style, explanation detail,
or formatting.
4. If the final results match or are mathematically
equivalent → score 1.
5. If the final results do not match → score 0.
6. Output ONLY a JSON object in this exact format:
{"score": 0 or 1, "explanation": "short explanation of
the comparison"}
[BEGIN DATA]
************
[Golden Answer]: O preço dos $72$ pintainhos será
representado por (...), pelo que cada pintainho foi
vendido a $511$ escudos.
************
[Model Answer]: Para resolver o problema, precisamos
(...) Conclusão, o preço por pintainho no ano anterior
era: $$\boxed{511}$$
************
[END DATA]
Provide your judgment now.
```

Figure 3: Example of a prompt used for judging a model's answer to an open-ended question

prompt template. For multiple-choice questions, the model's raw text output is parsed to extract the content inside the final `\boxed{}` expression. A question is counted as correctly answered if and only if the extracted letter exactly matches the ground-truth answer key. For open-ended questions, we use Kimi K2 Thinking (Team et al., 2026) as an LLM judge to check if model answers are correct, i.e., mathematically equivalent to the golden answers. An example of a prompt sent to a judge is shown in Figure 3. Accuracy is computed separately for each competition level.

The evaluated models include both closed models (GPT-5, Gemini 2.5 Flash, Claude Haiku 4.5) as well as open models (Qwen3, Gemma3, LLaMa3, DeepSeek Chat V3.1), covering a wide spectrum of model sizes and reasoning capabilities.

4 Analysis

The results in Table 2 show consistent performance patterns across models, levels, and question types. Frontier models such as GPT-5, Qwen3-235B, Gemini 2.5 Flash, and Claude Haiku 4.5 outperform all mid-sized and open-weight models. GPT-5 is the strongest model overall, achieving the highest accuracy across nearly every setting in both pt-PT and pt-BR. Open-source models exhibit wider variability, with the Qwen family showing notably strong scaling, while Gemma-3

Model	Level	pt-PT		pt-BR	
		MC	OE	MC	OE
GPT-5	L1	**96.15**	**100.00**	N/A	N/A
	L2	**94.67**	**89.58**	**90.68**	**90.62**
	L3	**96.50**	**92.75**	**92.86**	91.38
	L4	**96.55**	**85.52**	**93.33**	**89.47**
	L5	N/A	N/A	**90.00**	N/A
Qwen3-235B-A22B	L1	91.03	88.24	N/A	N/A
	L2	86.98	79.17	88.98	84.38
	L3	88.11	81.35	**92.86**	**93.10**
	L4	93.10	78.73	91.48	86.84
	L5	N/A	N/A	89.09	N/A
Llama-3.3-70b-Instruct	L1	41.03	38.24	N/A	N/A
	L2	32.54	17.71	35.59	34.38
	L3	23.78	18.13	35.00	34.48
	L4	24.14	20.36	32.96	18.42
	L5	N/A	N/A	31.82	N/A
Gemma-3-27b-it	L1	57.69	58.82	N/A	N/A
	L2	49.11	30.21	66.95	65.62
	L3	48.95	33.16	69.29	67.24
	L4	58.62	27.60	64.81	52.63
	L5	N/A	N/A	61.82	N/A
Deepseek-chat-v3.1	L1	87.18	73.53	N/A	N/A
	L2	84.02	67.71	83.05	78.12
	L3	88.81	72.02	91.43	86.21
	L4	79.31	66.52	89.26	81.58
	L5	N/A	N/A	86.36	N/A
Claude-Haiku-4.5	L1	80.77	67.65	N/A	N/A
	L2	74.56	47.92	83.90	68.75
	L3	76.92	56.48	84.29	70.69
	L4	68.97	47.51	81.48	60.53
	L5	N/A	N/A	81.82	N/A
Gemini-2.5-Flash	L1	83.33	88.24	N/A	N/A
	L2	75.74	72.92	83.90	84.38
	L3	83.92	74.61	88.57	91.38
	L4	79.31	71.04	85.93	71.05
	L5	N/A	N/A	89.09	N/A

Table 2: Performance of several models in multiple choice and open-ended questions and different levels.

and Llama-3.3 underperform significantly in mathematical reasoning. A consistent gap emerges between multiple-choice (MC) and open-ended (OE) performance. In general, all models achieve higher accuracy on MC questions, and the difference widens substantially at higher difficulty levels, often exceeding 10–20 points. GPT–5 shows the smallest MC–OE gap, reflecting more stable reasoning capabilities, whereas mid-range models display large drops, especially in pt-PT.

Table 3 shows that problems involving figures introduce a prominent difficulty spike. Across all models, performance deteriorates sharply on pt-PT items referencing figure code, with drops of 10–20 points even for frontier models and up to 56 points for weaker baselines. This degradation is more pronounced in OE questions, indicating that multimodal reasoning remains a key bottleneck despite the textual representation of figures. For weaker models, accuracy decreases systematically from Level 1 to Level 4 and 5. While GPT-5 and Qwen3-235B remain robust even at the highest difficulty

Model	Level	MC		OE	
		No Fig.	Fig.	No Fig.	Fig.
GPT-5	L1	**100.00**	**88.00**	**100.00**	**100.00**
	L2	**96.52**	**90.74**	**88.46**	**90.91**
	L3	**97.92**	**93.62**	**94.35**	**89.86**
	L4	**100.00**	**87.50**	**83.83**	**90.74**
Qwen3-235B-A22B	L1	98.11	76.00	**100.00**	73.33
	L2	90.43	79.63	82.69	75.00
	L3	95.83	72.34	86.29	72.46
	L4	95.24	**87.50**	80.24	74.07
Llama-3.3-70b-Instruct	L1	49.06	24.00	63.16	6.67
	L2	40.00	16.67	21.15	13.64
	L3	32.29	6.38	21.77	11.59
	L4	28.57	12.50	19.76	22.22
Gemma-3-27b-it	L1	66.04	40.00	73.68	40.00
	L2	59.13	27.78	42.31	15.91
	L3	58.33	29.79	33.87	31.88
	L4	66.67	37.50	25.75	33.33
Deepseek-chat-v3.1	L1	98.11	64.00	**100.00**	40.00
	L2	94.78	61.11	73.08	61.36
	L3	94.79	76.60	76.61	63.77
	L4	80.95	75.00	68.26	61.11
Claude-Haiku-4.5	L1	94.34	52.00	89.47	40.00
	L2	84.35	53.70	57.69	36.36
	L3	83.33	63.83	61.29	47.83
	L4	76.19	50.00	47.90	46.30
Gemini-2.5-Flash	L1	94.34	60.00	94.74	80.00
	L2	84.35	57.41	78.85	65.91
	L3	86.46	78.72	83.06	59.42
	L4	80.95	75.00	71.86	68.52

Table 3: Impact of visual content (figures) in pt-PT.

levels, other models exhibit sharp declines.

Some scaling trends are also noticeable for Qwen3, as shown in Table 4. Performance increases smoothly from 8B to 14B, 32B, 30B-A3B, and 235B in nearly all dimensions of the table. Scaling benefits are especially pronounced in OE tasks, indicating that larger models gain disproportionately in reasoning-heavy settings. Notably, Qwen3-14B performs unexpectedly well on pt-BR MC tasks, sometimes reaching accuracy levels comparable to older proprietary frontier models.

Comparing model architectures, Qwen3-235B is the strongest open-weight model exhibiting performances on par with the closed models. DeepSeek-Chat-v3.1 performs competitively on MC but struggles with OE and figure-based items. Llama-3.3-70B and Gemma-3-27b-it show limited mathematical reasoning capabilities.

5 Conclusion

We introduced MATH-PT, the first natively curated benchmark for mathematical reasoning in European and Brazilian Portuguese, sourced from real Olympiads and national exams. We observe that frontier models perform strongly on multiple-choice questions, with substantial drop in performance for open-ended questions and those those in-

Model	Level	pt-PT		pt-BR	
		MC	OE	MC	OE
Qwen3-8B	L1	84.62	82.35	N/A	N/A
	L2	76.33	58.33	87.29	75.00
	L3	78.32	61.14	87.14	72.41
	L4	65.52	50.23	81.48	52.63
	L5	N/A	N/A	76.36	N/A
Qwen3-14B	L1	88.46	82.35	N/A	N/A
	L2	**89.35**	73.96	**90.68**	81.25
	L3	86.71	76.68	90.00	84.48
	L4	82.76	71.04	88.15	78.95
	L5	N/A	N/A	**90.91**	N/A
Qwen3-32B	L1	88.46	79.41	N/A	N/A
	L2	83.43	70.83	88.98	84.38
	L3	85.31	73.06	89.29	82.76
	L4	**93.10**	65.16	85.19	76.32
	L5	N/A	N/A	87.27	N/A
Qwen3-30B-A3B	L1	82.05	82.35	N/A	N/A
	L2	82.84	64.58	82.20	75.00
	L3	86.71	63.73	89.29	74.14
	L4	82.76	55.20	84.44	50.00
	L5	N/A	N/A	80.00	N/A
Qwen3-235B-A22B	L1	**91.03**	**88.24**	N/A	N/A
	L2	86.98	**79.17**	88.98	**84.38**
	L3	**88.11**	**81.35**	**92.86**	**93.10**
	L4	**93.10**	**78.73**	**91.48**	**86.84**
	L5	N/A	N/A	89.09	N/A

Table 4: Impact of model size within the Qwen3 family.

volving figures. By releasing MATH-PT we aim to support reproducible evaluation and to encourage progress toward more capable Portuguese-based mathematical reasoning models.

Acknowledgments

We would like to thank José Mourão and António Salgueiro for their help in obtaining the LaTeX source for the OPM exams and solutions. 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), by the AMALIA project under Measure RE-C05-i08, and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações. ES acknowledge the support of "la Caixa" Foundation's LCF/BQ/PI22/11910028 award, as well as funds by MICIU/AEI/10.13039/501100011033 and the BERC 2022-2025 program funded by the Basque Government. ES research work is funded by Portuguese national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. DM, ACE, and JB acknowledge the support by the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) (SEI-260003/020694/2025) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (305692/2025-9).

References

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating LLMs on uncontaminated math competitions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Blanca Calvo Figueras, Eneko Sagarzazu, Julen Etxaniz, Jeremy Barnes, Pablo Gamallo, Iria de Dios-Flores, and Rodrigo Agerri. 2025. Truth knows no language: Evaluating truthfulness beyond English. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31204–31218, Vienna, Austria. Association for Computational Linguistics.

Amanda da Silva Oliveira, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas, and Eduardo José da Silva Luz. 2024. Toxic speech detection in Portuguese: A comparative study of large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6884–6915, Bangkok, Thailand. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023.

Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Felipe Paula, Cassiana Roberta Lizzoni Michelin, and Viviane Moreira. 2024. Evaluation of question answer generation for Portuguese: Insights and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5315–5327, Miami, Florida, USA. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE. Association for Computational Linguistics.

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A. De Souza, Maria José Bocorny Finatto, and Aline Paes. 2025. Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477, Vienna, Austria. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14333–14368, Vienna, Austria. Association for Computational Linguistics.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China. Association for Computational Linguistics.

Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.