CROSS-CULTURAL AI REASONING: DEEPSEEK-R1'S MULTILINGUAL PERFORMANCE ON BRAZILIAN EDUCATIONAL ASSESSMENTS

ALEXANDRE THUROW BENDER[1], GABRIEL ALMEIDA GOMES[1],
ULISSES BRISOLARA CORRÊA[1], RICARDO MATSUMURA ARAUJO[1]

[1] *Universidade Federal de Pelotas - atbender@inf.ufpel.edu.br*
Universidade Federal de Pelotas - gagomes@inf.ufpel.edu.br
Universidade Federal de Pelotas - ulisses@inf.ufpel.edu.br
Universidade Federal de Pelotas - ricardo@inf.ufpel.edu.br

1. INTRODUCTION

Recent advancements in Large Language Models (LLMs) have transformed artificial intelligence capabilities, particularly in reasoning tasks (BROWN et al., 2020; BUBECK et al., 2023). Among these developments, DeepSeek-R1 represents a significant advancement in reasoning-focused architectures, designed to handle complex logical inference and multi-step problem-solving (AI, 2025). While these models demonstrate strong performance on international benchmarks, their effectiveness in culturally-specific educational contexts remains underexplored.

Brazil's National High School Exam (ENEM) offers a unique evaluation framework for AI reasoning capabilities. Established in 1998 and restructured in 2009, ENEM serves as Brazil's primary university admission mechanism through the Unified Selection System (SISU) (SANTOS, 2011). The exam's multidisciplinary structure, covering Natural Sciences, Human Sciences, Languages and Codes, and Mathematics, combined with its emphasis on Brazilian cultural contexts, makes it an ideal benchmark for assessing AI performance in diverse educational settings.

This paper presents the first systematic evaluation of DeepSeek-R1's performance on ENEM questions across three years (2022-2024), examining how this advanced reasoning LLM handles culturally-situated educational assessments. Using the pass@k metric, we assess performance across different subject domains and identify emergent behaviors such as self-translation capabilities when processing Brazilian Portuguese content.

2. METHODOLOGY

Our evaluation employs the ENEM dataset published by Maritaca AI on the Hugging Face platform (PIRES; AI, 2023; NUNES; AI, 2023), comprising questions from Brazil's National High School Exam across three years (2022-2024). ENEM questions are written in Brazilian Portuguese and require interdisciplinary reasoning, making them ideal for assessing AI capabilities in culturally-specific educational contexts.

We evaluated DeepSeek-R1 (32B version with 4-bit quantization) using the Pass@K metric, which addresses LLM output variability by allowing multiple sampling attempts per question. For each question, we generated 10 independent responses with identical prompts but varying sampling parameters.

The Pass@K probability for each problem i, with n_i independent solutions generated and c_i correct solutions, is computed as:

This study was financed in part by CAPES – Finance Code 001, FAPERGS Award Agreement 22/2551-0000598-5.W, and NVIDIA Corporation with the donation of the Titan X Pascal GPU.

$$\text{Pass@k}(n_i, c_i, k) = 1 - \frac{\binom{n_i - c_i}{k}}{\binom{n_i}{k}}$$

For computational efficiency, we implemented the unbiased estimator:

$$\text{Pass@k}(n_i, c_i, k) = \begin{cases} 1.0 & \text{if } n_i - c_i < k \\ 1 - \prod_{j=n_i-c_i+1}^{n_i} \left(1 - \frac{k}{j}\right) & \text{otherwise} \end{cases}$$

The overall metric averages across all problems:

$$\text{Pass@k} = \frac{1}{|P|} \sum_{i \in P} \text{Pass@k}(n_i, c_i, k)$$

We constructed a system prompt that grounds the model in answering Brazilian educational assessment questions, with clear instructions to analyze each question, consider all alternatives, and provide structured responses that can be programmatically parsed. Each question included the question text and all answer alternatives (A, B, C, D, or E), with the model generating explicit reasoning steps followed by a clearly indicated final answer selection. Each prompt is run $n = 10$ times.

3. RESULTS AND DISCUSSION

Our evaluation reveals significant patterns in DeepSeek-R1's performance across different years and subject areas. Table 1 presents the overall annual performance, showing consistent improvement in pass@k values from 2022 to 2024 across all k values. Most notably, pass@1 increased from 0.825 in 2022 to 0.907 in 2024, representing approximately a 10% improvement over this three-year period.

Year	Pass@1	Pass@2	Pass@3	Pass@4	Pass@5
2022	0.825	0.875	0.894	0.905	0.911
2023	0.868	0.905	0.919	0.928	0.935
2024	0.907	0.936	0.944	0.948	0.950

Table 1: Pass@K Metrics for Different Years

The breakdown by subject area in Table 2 provides deeper insights into performance variations across disciplines. Human Sciences consistently demonstrates the highest performance across all years, achieving near-perfect results in 2023 and 2024 with pass@1 values of 0.990 and 0.988 respectively, and perfect scores (1.000) for higher k values.

Mathematics presents an interesting case, showing the lowest performance in 2022 and 2023 (pass@1 of 0.730 and 0.620 respectively), but experiencing the most dramatic improvement in 2024 with a pass@1 value of 0.885. This substantial gain from 2023 to 2024 represents the largest year-over-year improvement observed in any subject area.

Languages and Codes maintained relatively stable performance across the years, with pass@1 values ranging from 0.862 to 0.895. Similarly, Natural Sciences

Year	Area	pass@1	pass@2	pass@3	pass@4	pass@5
2022	Human Sciences	0.922	0.933	0.938	0.941	0.944
	Languages and Codes	0.867	0.927	0.952	0.963	0.969
	Mathematics	0.730	0.799	0.822	0.836	0.846
	Natural Sciences	0.763	0.827	0.852	0.866	0.875
2023	Human Sciences	0.990	0.999	1.000	1.000	1.000
	Languages and Codes	0.895	0.921	0.932	0.938	0.942
	Mathematics	0.620	0.700	0.737	0.759	0.776
	Natural Sciences	0.881	0.930	0.948	0.960	0.970
2024	Human Sciences	0.988	0.999	1.000	1.000	1.000
	Languages and Codes	0.862	0.900	0.911	0.916	0.918
	Mathematics	0.885	0.920	0.925	0.926	0.926
	Natural Sciences	0.844	0.888	0.914	0.930	0.938

Table 2: Pass@k Evaluation Results by Year and Subject Area

showed steady improvement. The consistency in these domains contrasts with the volatility observed in Mathematics.

Our examination of the model's reasoning process reveals behavioral patterns that may contribute to the performance trends. Of particular interest is the model's approach to non-English questions, exemplified by its handling of Portuguese content. Despite receiving no explicit translation instructions, the model demonstrated sophisticated self-translation capabilities, first translating Portuguese text internally before proceeding with analysis. This emergent behavior likely contributes to the exceptional performance in Human Sciences, where many questions involve cross-cultural and multilingual understanding.

4. CONCLUSION

Our evaluation of DeepSeek-R1 on Brazil's ENEM examination reveals strong reasoning capabilities within culturally-specific educational contexts. Results show impressive performance, particularly in Human Sciences with pass@1 rates exceeding 98% in 2023-2024, suggesting that recent advances in AI reasoning extend effectively beyond typical AI benchmarks to diverse cultural contexts.

The model's self-translation behavior and structured reasoning approach likely contribute to its performance in Human Sciences, allowing navigation of linguistically complex content without explicit translation instructions. Meanwhile, Mathematics exhibited the most variable performance across years, with notably stronger results in 2024 contrasting with lower scores in 2022 and 2023.

Despite promising results, limitations persist: performance variations across domains and years indicate inconsistencies, particularly in mathematical reasoning tasks. Our evaluation focused primarily on multiple-choice questions and may not fully capture performance on open-ended problem-solving that requires extended reasoning.

Future research should investigate whether self-translation behavior varies across different domains and languages, and whether this capability was explicitly built into the model's training or emerged as an emergent property. As AI technologies become increasingly integrated into global educational environments, understanding performance on culturally-specific assessments is crucial for ensuring equitable implementation while addressing remaining inconsistencies across knowledge domains.

REFERENCES

AI, D. Deepseek-r1: Advancing reasoning capabilities in large language models. **arXiv preprint arXiv:2401.14196**, 2025.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

BUBECK, S.; CHANDRASEKARAN, V.; ELDAN, R.; GEHRKE, J.; HORVITZ, E.; KAMAR, E.; LEE, P.; LEE, Y. T.; LI, Y.; LUNDBERG, S. et al. Sparks of artificial general intelligence: Early experiments with gpt-4. **arXiv preprint arXiv:2303.12712**, 2023.

NUNES, M.; AI, M. Enem questions: Brazilian educational assessment dataset. **Hugging Face Dataset**, 2023.

PIRES, J.; AI, M. Enem dataset: A comprehensive collection of brazilian national high school exam questions. **Hugging Face Dataset**, 2023.

SANTOS, C. A. d. O enem e o sisu: democratização do acesso ao ensino superior. **Educação em Revista**, v. 27, n. 1, p. 45–68, 2011.