

# AN EVALUATION OF LARGE LANGUAGE MODELS FOR SOLVING MATH PROBLEMS IN ALBANIAN LANGUAGE

Vikensa Grabocka, Marjana Prifti Skënduli



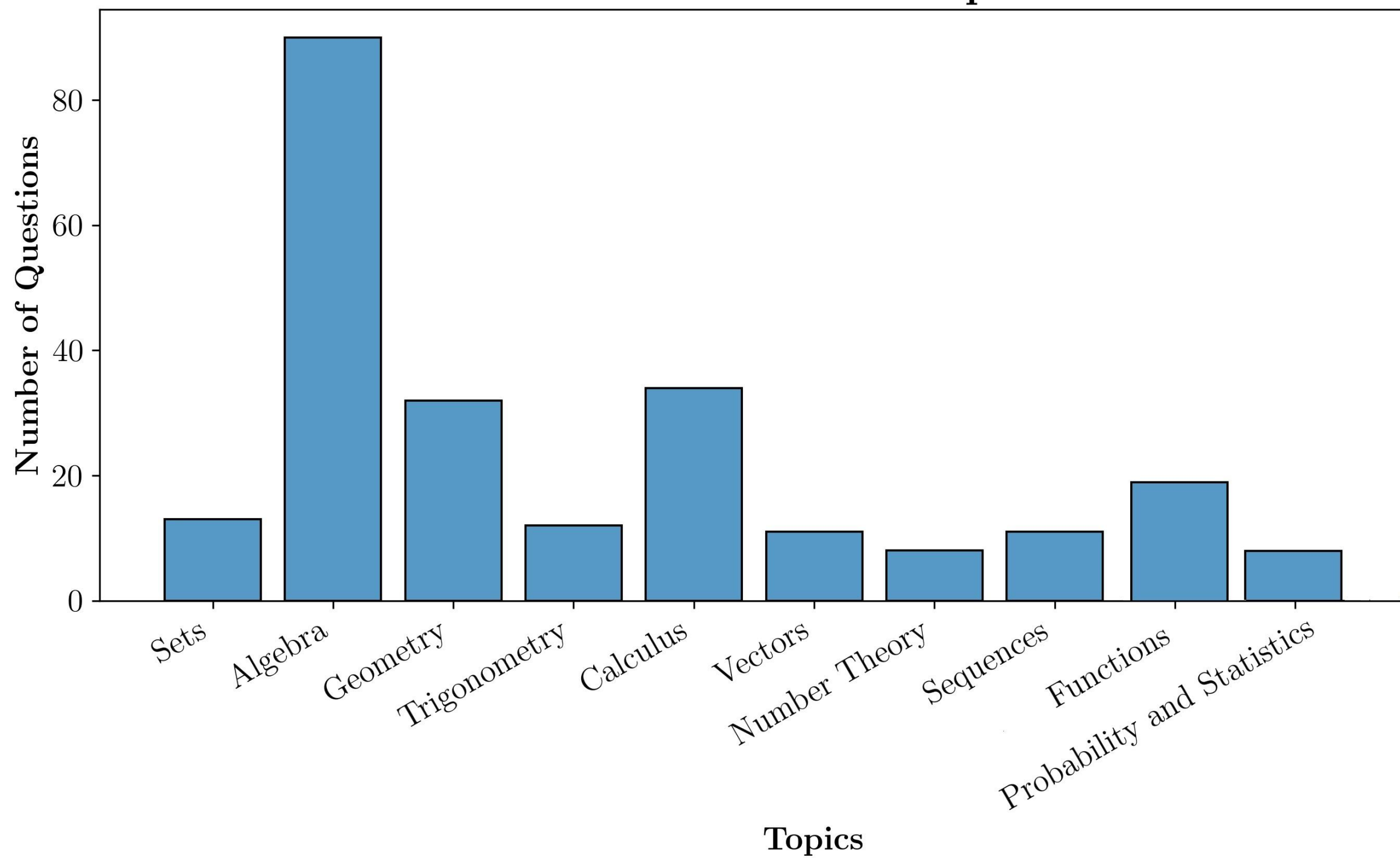
## INTRODUCTION

- Large Language Models (LLMs) represent a class of machine language models designed to process natural language tasks.
- The performance of 13 LLMs was tested: DeepSeek R1 and DeepSeek V3, o1 mini, GPT 4.1 mini, GPT 4.1, GPT 4o, GPT 4o mini, GPT 3.5 Turbo, Chat-GPT 4o from OpenAI and Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini 2.0 Flash, Gemini 2.5 Pro from Google .

## DATASET

- 238 multiple choice questions from the final, mandatory high-school examinations in mathematics in Albania.

Distribution of Topics

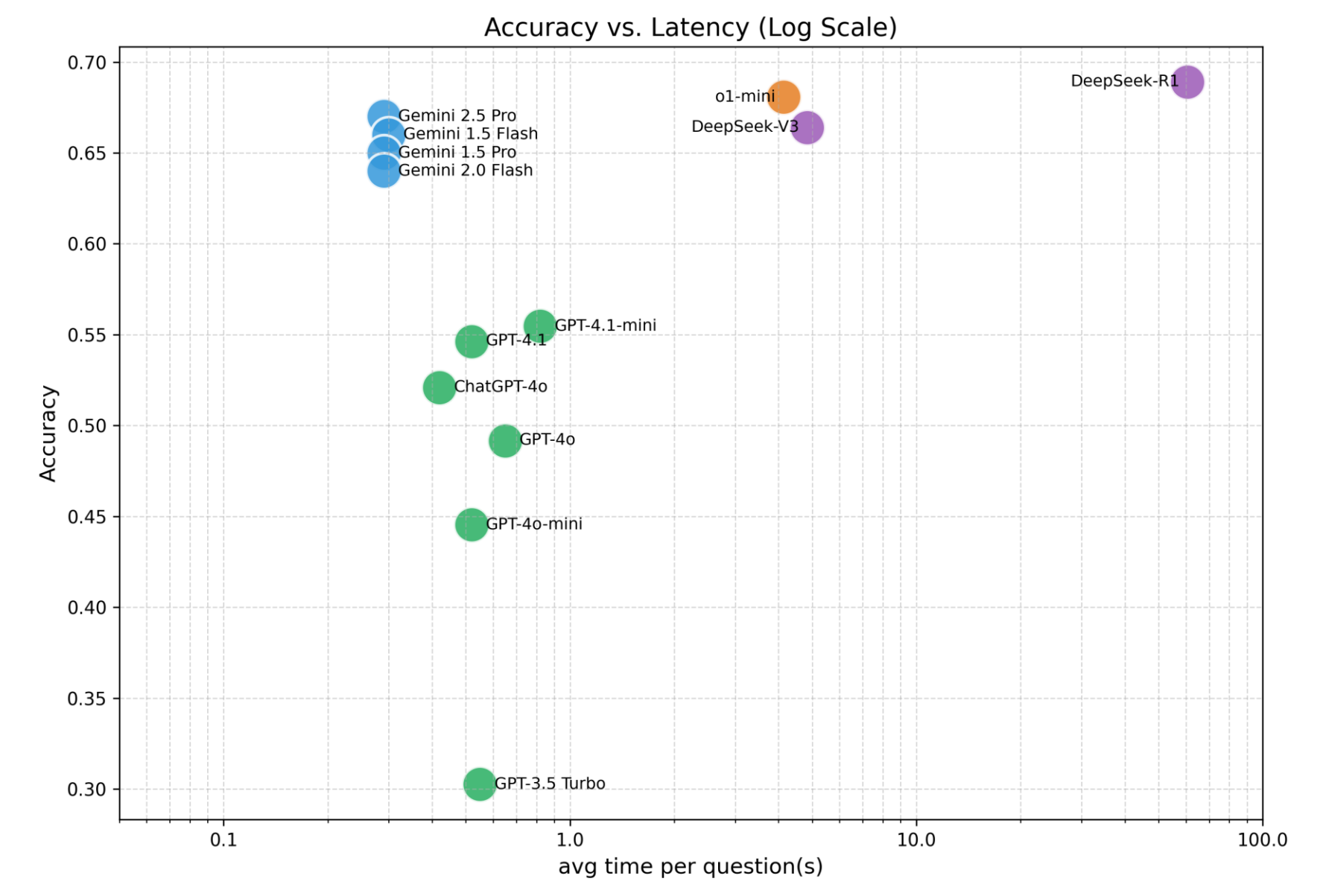


## METHODOLOGY

- The corresponding APIs for each model were used.
- The model was instructed in the prompt to return explicitly the correct alternative only.
- The answer and the time it took the model to solve the problem was recorded.

## RESULTS

- Three best performing models are the reasoning models.
- DeepSeek R1 achieves the best performance with an accuracy of 68.91% followed by o1 mini with 68.07% and Gemini 2.5 Pro with 66.81%.



ACCURACY PER MODEL PER TOPIC

Model	Algebra	Calculus	Functions	Geometry	Number Theory	Probability & Statistics	Sequences	Sets	Trigonometry	Vectors
GPT-4o-mini	52.2	35.3	31.6	37.5	37.5	37.5	27.3	69.2	58.3	36.4
GPT-4.1-mini	54.4	50.0	57.9	56.2	37.5	50.0	45.5	61.5	66.7	72.7
GPT-4.1	54.4	50.0	68.4	53.1	50.0	50.0	36.4	69.2	58.3	54.5
GPT-3.5-Turbo	25.6	29.4	42.1	21.9	37.5	37.5	27.3	46.2	25.0	27.3
GPT-4o	54.4	38.2	52.6	46.9	37.5	50.0	27.3	53.8	66.7	36.4
Deepseek V3	63.3	64.7	78.9	65.6	62.5	62.5	63.6	61.5	75.0	72.7
Deepseek-R1	67.8	61.8	78.9	62.5	75.0	87.5	54.5	84.6	66.7	72.7
ChatGPT-4o	55.6	38.2	47.4	53.1	62.5	62.5	36.4	61.5	66.7	45.5
o1-mini	70.0	61.8	78.9	56.2	75.0	75.0	54.5	76.9	75.0	63.6
Gemini-2.0-Flash	62.2	58.8	78.9	65.6	50.0	87.5	54.5	76.9	66.7	72.7
Gemini-1.5-Flash	62.2	64.7	73.7	65.6	50.0	87.5	63.6	76.9	75.0	72.7
Gemini-1.5-Pro	61.1	58.8	73.7	65.6	62.5	87.5	63.6	76.9	66.7	72.7
Gemini-2.5-Pro	63.3	64.7	78.9	65.6	50.0	87.5	63.6	76.9	66.7	72.7

AVERAGE RESPONSE TIME

Model	Algebra	Calculus	Functions	Geometry	Number Theory	Probability & Statistics	Sequences	Sets	Trigonometry	Vectors
GPT-4o-mini	0.59	0.49	0.46	0.48	0.53	0.56	0.44	0.48	0.47	0.42
GPT-4.1-mini	0.82	0.85	0.73	0.87	0.71	0.74	0.79	1.06	0.72	0.74
GPT-4.1	0.54	0.43	0.64	0.57	0.61	0.49	0.41	0.47	0.46	0.38
GPT-3.5-Turbo	0.42	0.40	0.40	0.43	0.36	0.33	0.58	0.46	0.37	0.38
GPT-4o	0.66	0.61	0.61	0.60	0.58	1.10	0.58	0.62	0.90	0.61
Deepseek V3	4.81	4.82	4.73	4.88	4.95	4.54	4.80	4.94	4.91	5.21
Deepseek-R1	63.64	75.29	34.53	62.76	59.16	37.42	94.05	60.74	47.35	28.46
ChatGPT-4o	0.51	0.55	0.49	0.56	0.50	0.54	0.59	0.84	0.54	0.55
o1-mini	4.49	3.56	3.61	3.81	3.81	3.73	6.96	3.99	3.42	3.30
Gemini-2.0-Flash	0.30	0.29	0.28	0.29	0.31	0.34	0.30	0.28	0.29	0.27
Gemini-1.5-Flash	0.29	0.30	0.28	0.28	0.32	0.29	0.27	0.35	0.30	0.33
Gemini-1.5-Pro	0.28	0.29	0.27	0.29	0.28	0.30	0.30	0.28	0.30	0.32
Gemini-2.5-Pro	0.28	0.29	0.29	0.30	0.27	0.30	0.29	0.29	0.32	0.30

## CONCLUSION

- Even the best models don't surpass 70%, showing limited reliability in other languages, apart from English.
- Performance is higher on Math topics such as "Sets" and "Probability & Statistics", but lower on "Sequences".

## References

1. Google DeepMind. 2023. Introducing Gemini: our largest and most capable AI model. <https://blog.google/technology/ai/google-gemini-ai/>
2. DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. doi:10.48550/arXiv.2501.12948 arXiv:2501.12948
3. DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. doi:10.48550/arXiv.2412.19437 arXiv:2412.19437
4. OpenAI. 2025. Models - OpenAI API. <https://platform.openai.com>
5. Siddhartha Prasad, Ben Greenman, Tim Nelson, and Shirram Krishnamurthi. 2023. Generating Programs Trivially: Student Use of Large Language Models. In Proceedings of the ACM Conference on Global Computing Education Vol 1 (Hyderabad, India) (CompEd 2023). Association for Computing Machinery, New York, NY, USA, 126–132. doi:10.1145/3576882.3617921



12<sup>th</sup> ACM Celebration of Women in Computing: womENCourage™  
 Braşov, Romania  
 17-19 September, 2025  
 Theme: Computer Science: a Catalyst for Educational Change

