SOVEREIGN: Gemini evaluation benchmark results MMLU HumanEval GSM8K MATH performance scores Google

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

We introduce ChatGLM, an evolving family of large language models that we have been developing over time. This report primarily focuses on the GLM-4 language series, which includes GLM-4, GLM-4-Air, and GLM-4-9B. They represent our most capable models that are trained with all the insights and lessons gained from the preceding three generations of ChatGLM. To date, the GLM-4 models are pre-trained on ten trillions of tokens mostly in Chinese and English, along with a small set of corpus from 24 languages, and aligned primarily for Chinese and English usage. The high-quality alignment is achiev

1 Introduction

Analysis of: ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. Research goal: Gemini evaluation benchmark results MMLU HumanEval GSM8K MATH performance scores Google.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

13 papers retrieved. 11 claims extracted, 11 verified. Tribunal: 9.2/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
ChatGLM is an evolving family of large language models developed over time.	✓	0.20
The GLM-4 language series includes GLM-4, GLM-4-Air, and GLM-4-9B.	✓	0.23
The GLM-4 models are pre-trained on ten trillions of tokens mostly in Chinese and English, along with a small set of cor	✓	0.31
The high-quality alignment of GLM-4 models is achieved via a multi-stage post-training process, which involves supervise	✓	0.29
GLM-4 closely rivals or outperforms GPT-4 in terms of general metrics such as MMLU, GSM8K, MATH, BBH, GPQA, and HumanEva	✓	0.32
GLM-4 gets close to GPT-4-Turbo in instruction following as measured by IFEval.	✓	0.24
GLM-4 matches GPT-4 Turbo (128K) and Claude 3 for long context tasks.	✓	0.23
GLM-4 outperforms GPT-4 in Chinese alignments as measured by AlignBench.	✓	0.23
The GLM-4 All Tools model is aligned to understand user intent and autonomously decide when and which tool(s) to use.	✓	0.27
The GLM-4 All Tools model can use tools such as web browser, Python interpreter, text-to-image model, and user-defined f	✓	0.28
In practical applications, the GLM-4 All Tools model matches and even surpasses GPT-4 All Tools in tasks like accessing	✓	0.35

References

- https://doi.org/10.48550/arxiv.2406.12793
- https://doi.org/10.1007/s11704-026-60308-3

- https://doi.org/10.36227/techrxiv.170956672.21573677/v1