SOVEREIGN: What is the impact of context length on the performance of Mixtral 8x7B versus single-check 7B models on the M

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Language models demonstrate both quantitative improvement and new qualitative capabilities with increasing scale. Despite their potentially transformative impact, these new capabilities are as yet poorly characterized. In order to inform future research, prepare for disruptive new model capabilities, and ameliorate socially harmful effects, it is vital that we understand the present and near-future capabilities and limitations of language models. To address this challenge, we introduce the Beyond the Imitation Game benchmark (BIG-bench). BIG-bench currently consists of 204 tasks, contributed b

1 Introduction

Analysis of: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Research goal: What is the impact of context length on the performance of Mixtral 8x7B versus single-check 7B models on the MMLU benchmark when evaluating long-context reasoning capabilities?.

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

10 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.8/10 → 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
BIG-bench currently consists of 204 tasks, contributed by 450 authors across 132 institutions	✓	0.28
BIG-bench focuses on tasks that are believed to be beyond the capabilities of current language models	✓	0.27
Model performance and calibration both improve with scale, but are poor in absolute terms	✓	0.23
A team of human expert raters performed all tasks in order to provide a strong baseline	✓	0.25
The benchmark consists of tasks from linguistics, childhood development, math, commonsense reasoning, biology, physics,	✓	0.27
Model performance is remarkably similar across model classes	✓	0.18
BIG-bench evaluates the behavior of OpenAI's GPT models, Google-internal dense transformer architectures, and Switch-sty	✓	0.29

References

- https://doi.org/10.48550/arxiv.2206.04615
- https://doi.org/10.1038/s41586-023-06291-2
- https://doi.org/10.1007/s11704-026-60308-3