SOVEREIGN: Does layer-wise score aggregation improve SuperGLUE task accuracy over last-layer baselines when evaluated on

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 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: Does layer-wise score aggregation improve SuperGLUE task accuracy over last-layer baselines when evaluated on out-of-distribution or adversarially constructed examples from the benchmark?.

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

9 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 9.0/10 $\rightarrow$ 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
Language models demonstrate both quantitative improvement and new qualitative capabilities with increasing scale.	✓	0.28
BIG-bench currently consists of 204 tasks, contributed by 450 authors across 132 institutions.	✓	0.28
Task topics in BIG-bench are diverse, drawing problems from linguistics, childhood development, math, common-sense reaso	✓	0.34
BIG-bench focuses on tasks that are believed to be beyond the capabilities of current language models.	✓	0.30
Model performance and calibration both improve with scale, but are poor in absolute terms (and when compared with rater	✓	0.26
Performance is remarkably similar across model classes, though with benefits from sparsity.	✓	0.21

References

- https://doi.org/10.1109/access.2024.3365742
- https://doi.org/10.48550/arxiv.2206.04615
- https://doi.org/10.1186/s40537-020-00392-9