Large Language Model Scale and Accuracy Degradation on Humanity Last Exam

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: What is the correlation between model parameter scale and accuracy degradation on the Humanity Last Exam subset for models exceeding 100B parameters. 6 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Using domain specific benchmarks for responsible LLM deployment. Research question: What is the correlation between model parameter scale and accuracy degradation on the Humanity Last Exam subset for models exceeding 100B parameters?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

1 papers retrieved. 6 claims extracted; 5 independently verified. Quality review score: 7.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Els Grans Models de Llenguatge (LLMs) han millorat l'eficincia, qualitat i escalabilitat dels processos en mltiples do	✓	0.24
El desplegament responsable de LLM presenta reptes relacionats amb el rendiment, el consum de recursos, la regulaci i l	✓	0.32
La selecci de models basada nicament en mtriques de capacitat o rendiment s insuficient per donar suport a una presa	✓	0.37
El projecte de recerca ML-Compass s un marc d'optimitzaci desenvolupat per donar suport a decisions de desplegament d'	✓	0.33
ML-Compass proporciona un marc unificador que formula la selecci com un problema d'optimitzaci amb restriccions, oferi	✓	0.42
HealthBench s un benchmark clnic de LLM d'ltima generaci.	×	0.14

References

- https://openalex.org/W7139610338