Language Model Perplexity and Downstream Reasoning Performance: A Multi-Study Synthesis

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the relationship between language model perplexity and downstream reasoning task performance v7. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation. Research question: What is the relationship between language model perplexity and downstream reasoning task performance v7.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

16 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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
Experiments were conducted on the Ollama platform using a Python script that automated the evaluation of 39 large langua	×	0.15
All tests were conducted with a fixed seed value of 42.	×	0.02
Prompt 1: gemma2:9b = 63 correct, llama3.1 = 47 correct	×	0.01
Prompt 2: gemma2:9b = 60 correct, llama3.1 = 33 correct	×	0.01
Prompt 3: gemma2:9b = 58 correct, llama3.1 = 37 correct	×	0.01
Prompt 4: gemma2:9b = 55 correct, llama3.1 = 36 correct	×	0.01
Prompt 5: gemma2:9b = 58 correct, llama3.1 = 42 correct	×	0.01
The evaluation results were published on the Hugging Face platform as three distinct datasets: AI Turkish MMLU Leaderboa	×	0.10
Table 1 summarizes the performance of selected models: gpt-4o (84.84% accuracy), claude-3.5 (84.40% accuracy), llama3.3:	×	0.05
Table 2 illustrates the performance of select models in key fields: gpt-4o (TUS: 91%, KPSS: 74.5%, Driver's License: 97%	×	0.01
A semantic similarity model, 'paraphrase-multilingual-mpnet-base-v2,' was employed to compute similarity scores between	×	0.05
Models with robust tokenization strategies tailored to Turkish morphology consistently outperformed others.	×	0.08
Fine-tuned models achieved substantial performance gains, though challenges such as catastrophic forgetting were observe	×	0.06

References

- http://arxiv.org/abs/2402.04177v3
- http://arxiv.org/abs/2501.00593v2
- http://arxiv.org/abs/2407.04973v1