SOVEREIGN: How does the number of iterative self-reflection steps affect the accuracy and inference efficiency of languag

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Small Language Models (SLMs) offer computational efficiency and accessibility, yet a systematic evaluation of their performance and environmental impact remains lacking. We introduce SLM-Bench, the first benchmark specifically designed to assess SLMs across multiple dimensions, including accuracy, computational efficiency, and sustainability metrics. SLM-Bench evaluates 15 SLMs on 9 NLP tasks using 23 datasets spanning 14 domains. The evaluation is conducted on 4 hardware configurations, providing a rigorous comparison of their effectiveness. Unlike prior benchmarks, SLM-Bench quantifies 11 me

1 Introduction

Analysis of: SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts–Extended Version. Research goal: How does the number of iterative self-reflection steps affect the accuracy and inference efficiency of language agents on the WebShop benchmark when using Reflexion versus standard fine-tuning approaches?.

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. 11 claims extracted, 2 verified. Tribunal: 2.5/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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
SLM-Bench evaluates 15 Small Language Models (SLMs) across 9 tasks using 23 datasets from 14 domains.	✓	0.30
SLM-Bench includes environmental impact metrics such as energy consumption and CO2 emissions.	×	0.14
SLM-Bench evaluates SLMs on 4 hardware configurations.	✓	0.21
GPT-Neo-1.3B has 1.37 F1 Score and 2.46 ROUGE score according to SLM-Bench.	×	0.05
Dolly-v2-3B was released in December 2022 and has 3 F1 Score and 5.8 ROUGE score.	×	0.01
LLaMA-2-7B was released in July 2023 and achieves 6.47 F1 Score and 13 ROUGE score.	×	0.01
TinyLlama-1.1B was released in August 2023 and has 1.1 F1 Score and 2 ROUGE score.	×	0.01
Mistral-7B was released in September 2023 and achieves 7 F1 Score and 13 ROUGE score.	×	0.01
Phi-1.5B was released in December 2023 and has 1.42 F1 Score and 2.84 ROUGE score.	×	0.01
SLM-Bench evaluates models on tasks including Open-domain Question Answering, Common Sense Reasoning, Mathematics Proble	×	0.07
SLM-Bench uses evaluation metrics including Accuracy, F1 Score, BLEU, ROUGE, METEOR, and Perplexity.	×	0.12

References

- http://arxiv.org/abs/2207.01206v4
- http://arxiv.org/abs/2508.15478v2
- http://arxiv.org/abs/2306.13030v2