

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Cognitively Grounded Benchmark Generation for Robust Compositional Reasoning in Large Language Models

Permalink

<https://escholarship.org/uc/item/30p7420x>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Author

Gomez-Ramirez, Danny A. J.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

MEMBER ABSTRACT FOR COGSCI2026

by

Danny Arlen de Jesús Gómez Ramírez

TITLE:

Cognitively Grounded Benchmark Generation for Robust Compositional Reasoning in Large Language Models

SHORT TITLE:

Cognitively Grounded Benchmarks for Compositional Reasoning in LLMs

ABSTRACT:

Large language models (LLMs) can write fluent mathematics, yet their reasoning collapses when proofs require verifiable composition across many dependent steps. We propose a cognitively grounded protocol for generating robust benchmark datasets in advanced mathematics by operationalizing seven human-inspired dimensions of expert practice: concept formation, dualization, negative knowledge, transfer, invariance control, lemma synthesis, and counterexample search. Rather than toy tasks, we construct meta-prompts from research-grade problem fragments and evaluate complete solution traces, auditing faithfulness and invariant preservation step by step. Under matched conditions, four state-of-the-art systems show a global breaking degree of more than 90% on stress tests, with failures concentrated in long-horizon planning, premise selection, lemma construction, and counterexample discovery. We conclude with dataset design principles that maximize diagnostic power and reproducibility, and with training levers—rationale SFT, process supervision with reward models, and stepwise preference learning—that directly target step-level correctness.