Self-Refine Iterative Method Improves CodeGen-2B Accuracy on HELM Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the Self-Refine iterative method impact the accuracy of codegen-2b on HELM language understanding tasks compared to single-pass generation. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Iterative Refinement Improves Compositional Image Generation. Research question: How does the Self-Refine iterative method impact the accuracy of codegen-2b on HELM language understanding tasks compared to single-pass generation?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

11 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
Qwen-Image, Gemini 2.5 Flash Image (NanoBanana), and GPT-Image-1 are state-of-the-art text-to-image model families.	×	0.05
Qwen-Image is open-sourced among the three models mentioned.	×	0.03
The models are evaluated across three prominent compositional generation benchmarks: ConceptMix, T2I-CompBench, and TIIF	×	0.10
ConceptMix measures a model's ability to bind multiple concept categories under increasing compositional complexity, ran	×	0.05
T2I-CompBench evaluates open-world compositionality, including attribute binding, object–object relationships, numeracy,	×	0.03
TIIF-Bench focuses on fine-grained instruction following across diverse scenarios such as 3D perspective, logical negati	×	0.04
The Visual Jenga scene decomposition benchmark tests a model's ability to progressively remove objects from a scene in a	×	0.10
The evaluation follows the respective protocols of each benchmark and uses a strong multimodal language model.	×	0.03
Text-to-image (T2I) models are trained on large-scale datasets of image–caption pairs that lack structured reasoning tra	×	0.08
T2I models do not inherently develop capabilities like self-correction or iterative refinement, instead relying on one-s	×	0.11
The framework integrates four components: a text-to-image (T2I) model, a vision-language model (VLM) critic, an image ed	×	0.14
The pipeline allows the model to iteratively refine its outputs rather than relying solely on a single forward pass.	×	0.03
Parallel sampling increases diversity but does not fundamentally change the underlying generation process, nor does it a	×	0.05
Parallel sampling struggles with complex compositional prompts.	×	0.14

References

- http://arxiv.org/abs/2601.15286v1
- http://arxiv.org/abs/2601.18577v2
- http://arxiv.org/abs/2207.08179v1