Integrative Decoding Scaling with Sampling Iterations on TruthfulQA Benchmarks

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does the effectiveness of Integrative Decoding's differentiable decoding loop scale with the number of sampling iterations when evaluated on multiple-choice and open-ended generation tasks in the. 16 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Local and Global Decoding in Text Generation. Research question: Does the effectiveness of Integrative Decoding's differentiable decoding loop scale with the number of sampling iterations when evaluated on multiple-choice and open-ended generation tasks in the TruthfulQA benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

12 papers retrieved. 16 claims extracted; 4 independently verified. Quality review score: 4.5/10.

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
Locally-normalised decoding algorithms generally perform better than globally-normalised ones as evaluated by MAUVE scor	×	0.11
Local normalisation leads to longer, less repetitive, and overall higher-quality text.	×	0.09
The distortion introduced by local decoding is an important component contributing to its performance.	×	0.14
A language model defines a probability distribution over the set of all finite strings.	×	0.09
Language models are generally defined autoregressively.	×	0.05
Sampling from a language model reduces to sampling iteratively from conditional distributions until an end-of-sequence s	×	0.06
A T-maxlength language model assigns zero probability to any string longer than T.	×	0.04
For any T-length prefix with non-zero probability, the probability of the end-of-sequence symbol is 1.	×	0.00
Local decoding algorithms use a pruning function to modify the LM's output distribution.	✓	0.16
The pruning function in local decoding assigns zero probability to subwords not in the retained subset.	×	0.06
Local decoding algorithms normalise the pruned distribution to obtain locally normalised distributions.	✓	0.17
Sampling from locally normalised distributions is straightforward and can be done iteratively.	×	0.07
Independent Metropolis-Hastings (IMH) algorithm allows approximate sampling from globally-normalised distributions witho	✓	0.26
The experiments compare locally and globally normalised versions of top-k and top-π decoding algorithms.	✓	0.19
The experiments use Pythia models ranging from 70m to 2.8b in size.	×	0.03
The experiments include 8 settings for each algorithm, with k spanning from 5 to 10,000 and π from 0.01 to 0.99.	×	0.02

References

- http://arxiv.org/abs/2410.10810v1
- http://arxiv.org/abs/1211.2960v1
- http://arxiv.org/abs/2511.02463v3