What is the correlation between repeated-training costs and score stability when scaling generative recommenda

Assignee Research

June 10, 2026

Abstract

Generative recommendation models can model user behavior as sequences of events and provide a shared backbone for multiple recommendation tasks. In production, however, pre-training gains do not automatically translate into downstream application improvements: task headroom, repeated-training cost, serving latency, and item freshness all affect transfer. We describe our experience scaling a generative recommender from 2M to 1B backbone parameters, excluding embedding and decoding layers, in a production-scale title recommendation setting. Across multiple downstream tasks, we observe task-depen

1 Introduction

This paper examines: Towards Generalizable and Efficient Large-Scale Generative Recommenders. Research question: What is the correlation between repeated-training costs and score stability when scaling generative recommendation backbones across different user behavior sequence lengths?.

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 3.0/10.

3 Results

11 papers retrieved. 17 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
The 1B-backbone model improves MRR by +22.5% for Task A compared to the 2M-backbone baseline.	×	0.04
The 1B-backbone model improves MRR by +11.3% for Task B compared to the 2M-backbone baseline.	×	0.05
The 1B-backbone model improves MRR by +7.4% for Task C compared to the 2M-backbone baseline.	×	0.05
The 1B-backbone model improves MRR by +28.1% for cold-start titles compared to the 2M-backbone baseline.	×	0.06
Downstream integrations of the 1B-backbone model have produced positive outcomes in multiple production A/B tests.	×	0.06
The 1B-backbone model excludes embedding and decoding layers from the size count.	×	0.11
The shadow-traffic evaluation involves 1M users.	×	0.01
The largest gain on cold-start titles supports the semantic-tower design in Section 6.	×	0.05
The scaling-law analysis shows less remaining headroom for easier, time-dependent tasks.	×	0.08
The 1B-backbone model transfers across downstream integrations.	×	0.05
The 2M-backbone baseline has a P0 value of 0.31.	×	0.02
The 2M-backbone baseline has a P0 value of 0.60.	×	0.02
The 2M-backbone baseline has a P0 value of 1.07.	×	0.02
The 1B-backbone model shows a +11.3% improvement in MRR for Task B.	×	0.04
The 1B-backbone model shows a +7.4% improvement in MRR for Task C.	×	0.05
The 1B-backbone model shows a +28.1% improvement in MRR for cold-start titles.	×	0.06
The 1B-backbone model shows a +22.5% improvement in MRR for Task A.	×	0.04

References

- http://arxiv.org/abs/2605.23312v1
- http://arxiv.org/abs/2511.00176v1
- http://arxiv.org/abs/2511.06077v3