To what extent do pre-training gains in large-scale generative recommenders degrade when transferred to downst

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: To what extent do pre-training gains in large-scale generative recommenders degrade when transferred to downstream tasks with limited task headroom?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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
The 1B-backbone model improves MRR across all reported slices: +22.5% for Task A, +11.3% for Task B, +7.4% for Task C, a	×	0.06
The largest gain on cold-start titles supports the semantic-tower design in Section 6.	×	0.05
The smaller gain on Task C is consistent with the scaling-law analysis showing less remaining headroom for easier, time-	×	0.06
Downstream integrations of the model have also produced positive outcomes in multiple production A/B tests.	×	0.06
The 1B-backbone model transfers across downstream integrations.	×	0.05
P0 = 0.31	×	0.00
P0 = 0.60	×	0.00
P0 = 1.07	×	0.00
The 1B-backbone model shows a +11.3% improvement in MRR for Task B.	×	0.05
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.	×	0.07
The 1B-backbone model shows a +22.5% improvement in MRR for Task A.	×	0.05

References

- http://arxiv.org/abs/2605.23312v1
- http://arxiv.org/abs/2204.12833v3
- http://arxiv.org/abs/2311.15317v5