

[← Go to KDD 2026 Research Track Cycle 2 homepage \(/group?id=KDD.org/2026/Research_Track_Cycle_2\)](#)

On the Memorization and Generalization of Generative Recommendation



Yijie Ding (/profile?id=~Yijie_Ding2), Zitian Guo (/profile?id=~Zitian_Guo1), Jiacheng Li (/profile?id=~Jiacheng_Li2), Letian Peng (/profile?id=~Letian_Peng1), Shuai Shao (/profile?id=~Shuai_Shao5), Wei Shao (/profile?id=~Wei_Shao11), Xiaoqiang Luo (/profile?id=~Xiaoqiang_Luo2), Luke Simon (/profile?id=~Luke_Simon1), Jingbo Shang (/profile?id=~Jingbo_Shang2), Julian McAuley (/profile?id=~Julian_McAuley1), Yupeng Hou (/profile?id=~Yupeng_Hou1)

Published: 15 May 2026, Last Modified: 15 May 2026 SIGKDD 2026 Research Track
 Research Track Cycle 2, Senior Area Chairs, Area Chairs, Reviewers, Authors
 Revisions (/revisions?id=OhWJ3XL3Qc) BibTeX
 CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

Keywords: Generative Recommendation, Memorization, Generalization

Abstract:

In this work, we aim to identify the types of data where semantic ID-based generative recommendation (GR) models perform better than traditional item ID-based models. We analyze model behavior through the lens of memorization (reusing item transition patterns seen during training) versus generalization (composing known patterns to predict unseen item transitions). To this end, we propose a series of criteria to categorize each data instance according to the capability required to make a correct prediction. Extensive experiments show that GR models perform better on instances that require generalization, whereas item ID-based models perform better when memorization dominates. To explain this divergence, we shift the analysis from the item level to the token level, demonstrating that item-level generalization often reduces to token-level memorization for GR models. Finally, we show that the two paradigms are complementary. We propose a simple memorization-aware indicator that adaptively combines them on a per-instance basis, leading to improved overall recommendation performance.

Corresponding Author: Yupeng Hou (/profile?id=~Yupeng_Hou1)

Resubmission Flag: No

Artifact Pledge: Yes

Submission Guidelines Blind: Yes

Submission Guidelines Format: Yes

Submission Guidelines Limit: Yes

Submission Guidelines Authorship: Yes

Student Author: Yes

Serve As Reviewer: Yupeng Hou (/profile?id=~Yupeng_Hou1)

Reviewer Qualification: N.A. - A qualified author is provided in the previous question

LLM Usage Description: LLMs are used to polish the language.

Information Sharing Consent: Yes

Submission Number: 3583

Discussion (?id=OhWJ3XL3Qc#discussion)

Filter by reply type... ▾

Filter by author... ▾

Search keywords...

Sort: Newest First



Everyone

Program Chairs

Submission3583 Authors

Submission3583...

13 / 13 replies shown

Submission3583 Area...

Submission3583...

Submission3583...

Submission3583...

Submission3583...

Submission3583...

**Paper Decision**

Decision by Program Chairs 15 May 2026, 19:15 (modified: 15 May 2026, 19:30) Program Chairs, Authors

Revisions (/revisions?id=6MfFXb4BKw)

Decision: Accept**Comment:**

AC: ~Chaokun_Wang1

SAC: ~Francesco_Gullo1

METAREVIEW: This paper proposes a systematic study comparing semantic ID-based generative recommendation models with traditional item ID-based models through the lens of memorization and generalization. Also, a framework is proposed to categorize data instances based on the item transition patterns they contain. All reviewers support the acceptance of this paper based on the paper and the authors' rebuttal. Also, the reviewers acknowledge the functional divergence between GR and traditional recommendation paradigms presented in the paper. Of course, some reviewers have concerns on the scalability of the proposed methods. After reading the paper and the discussion among the authors and the reviewers, I recommend that this paper is accepted. In addition, the authors should carefully incorporate the clarifications and improvements in the rebuttal into the revised version.

Official Review of Submission3583 by Reviewer xdQP

Official Review by Reviewer xdQP 24 Mar 2026, 07:28 (modified: 15 May 2026, 19:30)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer xdQP, Authors

Revisions (/revisions?id=J7a0IopRqf)

Paper Summary:

This paper presents a systematic analytical framework for understanding when semantic ID-based generative recommendation (GR) models outperform traditional item ID-based models, and vice versa. The authors formalize the distinction through the lens of memorization (predicting based on item transitions seen during training) and generalization (composing novel predictions from known transition patterns). A series of criteria — transitivity, symmetry, 2nd-order symmetry, and substitutability — are proposed to categorize each test instance. Experiments across multiple public benchmarks confirm that GR models (represented by TIGER) excel on generalization-related instances, while item ID-based models (represented by SASRec) dominate memorization-related ones. To explain this divergence, the authors further shift the analysis to the token level, demonstrating that item-level generalization in GR often reduces to token-level prefix memorization. Building on these findings, a memorization-aware adaptive ensemble is proposed that uses maximum softmax probability (MSP) of the ID-based model as an instance-specific indicator for weighting the two paradigms, yielding consistent improvements over both individual models and naive fixed-weight ensembles.

Paper Strengths:

- 1. Principled and interpretable analytical framework:** The paper identifies a meaningful gap in the generative recommendation literature: prior work explains GR's advantage via design choices (tokenization, learning objectives) but lacks a systematic analysis of which data instances each paradigm handles well. The proposed categorization scheme — grounded in item transition patterns rather than item-level cold-start proxies — is clearly motivated and formally defined. The distinction between 1-hop and multi-hop generalization types (transitivity, symmetry, 2nd-order symmetry, substitutability) provides a fine-grained vocabulary that can serve as a shared reference point for future work in this area.
- 2. Token-level analysis provides a compelling mechanistic explanation:** The bridging insight — that item-level generalization in GR models often reduces to token-level prefix memorization — is both technically non-trivial and practically useful. The empirical evidence demonstrating that token memorization support correlates with GR's generalization gain, while prefix-sharing dilutes item memorization for GR, provides a mechanistic explanation that is absent from existing literature. This contribution advances theoretical understanding of why GR and ID-based models behave differently, beyond surface-level performance comparisons.
- 3. Breadth of evaluation and actionable practical contribution:** The evaluation spans a diverse collection of public benchmarks covering different domains and interaction densities, lending generality to the findings. The practical contribution — the MSP-based adaptive ensemble — is deliberately lightweight (training-free, using an existing ID-based model's confidence) and is empirically validated for both its correlation with memorization categories and its downstream performance benefit. The availability of source code further strengthens the paper's reproducibility.

Paper Weaknesses:

- 1. The scope of generative recommendation models could be broadened:** The experimental analysis primarily focuses on TIGER as the representative GR model. It would be beneficial for the authors to discuss or at minimum acknowledge how the findings might extend — or not — to other prominent GR architectures such as ActionPiece (Hou et al., ICML 2025), LETTER/LRURec (end-to-end learnable tokenization, SIGIR 2025), or UniGRec (Yang et al., TMLR 2025), which employ substantially different tokenization strategies. Including even a brief discussion of whether the token-level memorization/generalization dynamics hold under these alternative tokenization paradigms would strengthen the paper's claimed generality. A single additional GR model in the analysis would make the findings considerably more convincing.
- 2. The adaptive ensemble's incremental improvement warrants additional discussion:** The memorization-aware adaptive ensemble consistently improves over the fixed-weight ensemble, but the margin is quite narrow on several datasets. The work could be strengthened by a more thorough exploration of what determines the magnitude of improvement — for instance, the relationship between the performance crossover strength (Figure 8) and the gain from adaptive weighting is qualitatively described but not quantitatively analyzed. Additionally, it might be worth clarifying whether the adaptive ensemble's benefit is primarily concentrated in the memorization or generalization subsets, which would directly connect the ensemble contribution back to the paper's core analytical claims.
- 3. Uncategorized instances merit deeper investigation:** Section 2.5 introduces "uncategorized" instances — those that fall outside both memorization and generalization categories under the chosen maximum hop count. These instances constitute a non-trivial fraction of several datasets (up to roughly 10% for some). The current treatment is brief. A more thorough exploration of what characterizes these instances — whether they represent truly out-of-distribution items, noise, or patterns requiring longer-range dependencies than the chosen hop limit — would help readers better understand the completeness and limitations of the proposed categorization framework. Clarifying the sensitivity of the overall conclusions to the choice of maximum hop count would also be valuable.
- 4. Presentation of the token-level analysis could be made more accessible:** The prefix n-gram memorization definition and the subsequent dilution analysis (Section 4) introduce several interrelated metrics (ϕ , ψ , token memorization support) in fairly rapid succession. A clearer explanation — perhaps a worked example or a unifying diagram — showing how item-level generalization, token-level memorization support, and the dilution effect interrelate would help readers better appreciate the technical depth of this section and improve the paper's overall accessibility.

Resubmission:

N/A

Questions And Suggestions For Rebuttal:

1. The analysis is anchored on TIGER as the sole representative GR model. Could the authors clarify whether the observed memorization/generalization split and the token-level reduction are specific to TIGER's RQ-VAE-based semantic IDs, or whether similar patterns hold for GR models using alternative tokenization methods (e.g., end-to-end learnable tokenization or context-sensitive tokenization)? Even an analysis on one additional GR model would substantially bolster the generalizability claim.
2. For the adaptive ensemble experiments, the improvement over the fixed-weight ensemble is consistent but modest on several benchmarks. I would expect the authors to provide a breakdown of the adaptive ensemble's gain on memorization vs. generalization subsets separately, to directly verify that the memorization-aware weighting is doing what is intended rather than simply acting as a well-tuned global weight.
3. Regarding the uncategorized instances: what is the impact on the paper's main claims if uncategorized instances are excluded from the "overall" performance metric, compared to including them? Could the authors clarify whether the conclusion that "GR models outperform on overall performance" holds primarily because of their advantage on generalization instances, or if uncategorized instances also contribute significantly to this overall advantage?

Relevance: 4: High - The work is relevant to the Research track of KDD and is of broad interest to the community

Novelty: 4: High - The paper offers groundbreaking and transformative ideas or approaches that substantially advance the field or open up entirely new areas of research. The level of innovation is high, leading to major advancements and potentially inspiring further research and development.

Technical Quality: 4: High - The paper exhibits a high level of technical quality with a rigorous and well-executed methodology and analysis. The results are highly reliable, well-supported, and thorough. The work demonstrates technical excellence and sets a high standard for quality in the field.

Presentation: 3: Moderate - The paper is organized and generally clear. The writing is mostly free of grammatical and typographical errors, making it easy to read. Figures and tables are effectively used to support the text. The presentation facilitates understanding and conveys the key points effectively.

Reproducibility: 4: High - The paper offers a comprehensive and precise description of the methods, data, and procedures. Supplementary materials, including datasets and code, are complete, well-documented, and easily accessible. Reproducing the results would be straightforward and require minimal additional effort, ensuring high reproducibility.

Reviewer Confidence: 4: High - The reviewer is an expert in the subject area and has extensive knowledge of the research methods and context of the paper. They are highly confident in their ability to provide an accurate and thorough assessment. Their evaluation is based on deep expertise and a comprehensive understanding of the work.

Ethics Review Flag: No

Ethics Review Description: N/A



Rebuttal by Authors

Rebuttal

by Authors (Yijie Ding (/profile?id=~Yijie_Ding2), Zitian Guo (/profile?id=~Zitian_Guo1), Jiacheng Li (/profile?id=~Jiacheng_Li2), Letian Peng (/profile?id=~Letian_Peng1), +7 more (/group/edit?id=KDD.org/2026/Research_Track_Cycle_2/Submission3583/Authors))

11 Apr 2026, 04:46 (modified: 11 Apr 2026, 04:57)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=WpPYS8ouKL)

Rebuttal:

1. Does the memorization/generalization split hold beyond TIGER?

We evaluated two additional models: PSID, a GR model that resolves SID collision using non nearest-neighbor assignment, and LRURec[1], an item ID-based model based on efficient recurrent modules. We observe that PSID shows a breakdown pattern similar to TIGER, while LRURec behaves similarly to SASRec. Moreover, PSID consistently generalizes better but memorizes worse than LRURec. These results suggest that our findings are not specific to specific tokenization or architecture, and can extend to other GR and item ID-based models as well

| Model(M/G) | Sports | Beauty | Sci | Music | Office | Steam | Yelp |
|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| LRURec | .2580/.0100 | .3474/.0142 | .2698/.0109 | .2121/.0142 | .2649/.0077 | .3978/.0108 | .3028/.0256 |
| PSID | .1635/.0175 | .2432/.0188 | .2143/.0162 | .1962/.0187 | .2682/.0157 | .3871/.0143 | .1303/.0201 |

[1] Linear Recurrent Units for Sequential Recommendation

2. Is adaptive ensemble improvement concentrated in memorization or generalization?

We provide a fine-grained breakdown of the adaptive vs. fixed ensemble on memorization and generalization subsets. Overall, the adaptive ensemble improves over the fixed-weight ensemble on most generalization subsets, with mixed effects on memorization. To verify that the weighting behaves as intended, we report the average ensemble weight (α) on both subsets. Results show that the system assigns higher weights to SASRec on memorization than generalization instances, which is aligned with our intended design and helps compensate for the shortcoming for TIGER on memorization.

| Metric(M/G) | Sports | Beauty | Sci | Music | Office | Steam | Yelp |
|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Fixed | .2684/.0180 | .3813/.0176 | .2664/.0182 | .2208/.0213 | .2862/.0155 | .3958/.0158 | .2447/.0268 |
| Adapt | .2651/.0188 | .3733/.0190 | .2671/.0183 | .2211/.0214 | .2899/.0156 | .3958/.0159 | .2424/.0270 |
| Mean α | .459/.426 | .536/.520 | .423/.400 | .401/.396 | .297/.256 | .439/.407 | .581/.559 |

3. Does GR advantage come from generalization or uncategorized instances?

GR's advantage over item ID-based models comes primarily from generalization-related instances rather than uncategorized ones. There are two main reasons. First, as shown in Table1, uncategorized instances account for less than 10% of the test data, whereas generalization-related instances typically account for around 80%. Second, both models achieve near-zero performance on the uncategorized subset, suggesting that neither model can reliably make correct predictions there

Official Review of Submission3583 by Reviewer L7yC

Official Review by Reviewer L7yC  21 Mar 2026, 02:32 (modified: 15 May 2026, 19:30)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer L7yC, Authors

 Revisions (/revisions?id=VrWuFhj7vS)

Paper Summary:

Overall

This is an analytical paper, not chasing state-of-the-art numbers, but instead trying to answer the question of why generative recommendation (GR) models work well.

The framework is rigorously designed, the experiments cover a wide range of settings, and the token-level interpretation offers genuine insight. The main weaknesses are some subjectivity in the definitions and a fairly limited contribution from the ensemble section.

Pros

- Valuable research question. A systematic comparison of GR vs. ID-based methods is something the community genuinely needs, and this kind of structured analysis is currently lacking.
- Rigorous framework design. The Memorization/Generalization taxonomy is clearly defined, formally specified, and mutually exclusive.
- Strong consistency across seven datasets, which lends credibility to the conclusions.

- The token-level mechanism analysis is a highlight. The insight that "item-level generalization reduces to token-level memorization" is deep and does a good job of explaining GR model behavior.

Cons

- The memorization definition only considers the last-1-hop transition. Equation (4) checks whether the transition $[i_{t-1} \rightarrow i_t]$ appeared in the training set, while completely ignoring the context of other items in the user's history. The same transition can be much harder or easier to predict depending on its surrounding sequence context. This definition reduces what is fundamentally a context-dependent problem into a binary pattern-matching check.
- The symmetry assumption deserves scrutiny. The definition treats the presence of $[i_t \rightarrow i_{t-1}]$ in training as grounds for "symmetry generalization" to $[i_{t-1} \rightarrow i_t]$. But item transitions in recommendation are inherently directional — "bought A then B" and "bought B then A" reflect very different user behaviors. Using a reverse transition as evidence for inferring the forward one lacks behavioral grounding.
- Only two models are benchmarked. TIGER and SASRec are reasonable representatives of their respective paradigms, but the GR space now has many variants (e.g., LETTER, LRURec). To what extent do the conclusions generalize to other GR models? This is not discussed at all.

Questions

1. The memorization definition only looks at the 1-hop transition $[i_{t-1} \rightarrow i_t]$, ignoring longer sequence context. If you extended this to consider co-occurrence patterns across the most recent K items, how much would the categorization results change?
2. The symmetry generalization assumption is that the presence of $[i_t \rightarrow i_{t-1}]$ in training allows the model to infer $[i_{t-1} \rightarrow i_t]$. Is there any experimental evidence that the model actually exploits this symmetry? Or is this purely a classification criterion that may not correspond to the model's actual inference mechanism?
3. The improvement of adaptive ensemble over fixed-weight ensemble in Table 3 is very small (under 1% relative gain). Is this difference statistically significant? Under what conditions would adaptive weighting offer a meaningful practical advantage over fixed weights?

Paper Strengths:

Refer to summary

Paper Weaknesses:

Refer to summary

Resubmission:

N

Questions And Suggestions For Rebuttal:

Refer to summary

Relevance: 3: Moderate - The work is somewhat relevant to the Research track of KDD and is of narrow interest to a sub-community

Novelty: 3: Moderate - The paper introduces a new and interesting idea or approach that adds value to the field. The contribution is original and represents an advancement of existing knowledge, demonstrating solid innovation and creativity.

Technical Quality: 3: Moderate - The paper demonstrates solid technical quality with a sound methodology and thorough analysis. The results are reliable and well-supported. There may be minor issues, but they do not significantly undermine the overall quality. The work is competently executed and meets acceptable standards.

Presentation: 3: Moderate - The paper is organized and generally clear. The writing is mostly free of grammatical and typographical errors, making it easy to read. Figures and tables are effectively used to support the text. The presentation facilitates understanding and conveys the key points effectively.

Reproducibility: 3: Moderate - The paper provides a clear and detailed description of the methods, data, and procedures used. Supplementary materials, such as datasets and code, are available and sufficiently documented. Reproducing the results would be feasible with the provided information, though some effort may still be required.

Reviewer Confidence: 3: Moderate - The reviewer has a good understanding of the subject area and is familiar with the research methods and context of the paper. They feel confident in their ability to accurately assess the quality and significance of the work. Their evaluation is based on a solid grasp of the content and context.

Ethics Review Flag: No**Ethics Review Description:** N

Rebuttal by Authors

Rebuttal

by Authors ([👤 Yijie Ding \(/profile?id=~Yijie_Ding2\)](/profile?id=~Yijie_Ding2), [Zitian Guo \(/profile?id=~Zitian_Guo1\)](/profile?id=~Zitian_Guo1), [Jiacheng Li \(/profile?id=~Jiacheng_Li2\)](/profile?id=~Jiacheng_Li2), [Letian Peng \(/profile?id=~Letian_Peng1\)](/profile?id=~Letian_Peng1), +7 more (/group/edit?id=KDD.org/2026/Research_Track_Cycle_2/Submission3583/Authors))

📅 11 Apr 2026, 04:47 (modified: 11 Apr 2026, 04:59)

👤 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

📄 Revisions (/revisions?id=KiIu1Csckg)

Rebuttal:

1. Extending the Memorization Definition to Multi-hop Co-occurrence Patterns

We define this multi-hop co-occurrence pattern as ‘substitutability’ (Sec 2.4), a form of generalization. Although this can be viewed as a multi-hop extension of memorization, we believe it requires the model to bypass unnecessary intermediate items and select the appropriate multi-hop transition for prediction, which goes beyond direct recall of a learned 1-hop transition and requires generalization.

2. Why use symmetry as a generalization criterion?

We define symmetry based on fundamental user behavior patterns, rather than tailoring it to specific model architectures. For example, flipping the order of user actions (e.g., watching Iron Man then Captain America) often yields equally valid recommendations; thus, it is crucial for a model to infer this symmetric relationship. Prior studies, such as CL4SRec [2], explicitly use "item reordering" for data augmentation, making the model robust to handle the ‘flexibility’ order of interactions in the real world recommendation. Please refer to our response to Reviewer pFMx for our overall design principles.

[2] Contrastive Learning for Sequential Recommendation. SIGIR 2021.

3. Is the adaptive ensemble improvement statistically significant, and when is it practically useful?

We thank the reviewer for the insightful question. We conducted paired t-tests on NDCG@10 scores between the fixed and adaptive ensembles, treating each user as a paired observation. The improvements are statistically significant at $\alpha=0.05$ on datasets including Office ($p<0.001$), Music ($p=0.0076$), Sports ($p=0.0213$), and Beauty ($p=0.0357$), while it degrades to the fixed-weight baseline on the remaining datasets.

Instead of proposing a universally superior method, our goal is to show these models are complementary. The gain of the adaptive ensemble is primarily determined by the amount of performance crossover (i.e., whether the models exhibit opposite performance gaps on memorization versus generalization subsets). When this crossover is minimal, the approach degrades to the fixed ensemble baseline. Ultimately, the practical advantage of the adaptive ensemble is achieving potential gains by complementing the strengths of both paradigms through a simple, training-free indicator. We will add an explicit discussion of these dynamics in the revised manuscript.

4. Do the conclusions generalize to other model variants?

Please see our response for reviewer xdQP.



Official Comment by Reviewer L7yC

Official Comment by Reviewer L7yC 📅 14 Apr 2026, 04:22

👤 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I have read the authors' response, and I have no further concerns. Positive score would be maintained.



➔ *Replying to Official Comment by Reviewer L7yC*

Official Comment by Authors

Official Comment

by Authors (Yijie Ding (/profile?id=~Yijie_Ding2), Zitian Guo (/profile?id=~Zitian_Guo1), Jiacheng Li (/profile?id=~Jiacheng_Li2), Letian Peng (/profile?id=~Letian_Peng1), +7 more (/group/edit?id=KDD.org/2026/Research_Track_Cycle_2/Submission3583/Authors))

18 Apr 2026, 01:40

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We are glad that our responses have addressed your concerns. Once again, we sincerely appreciate your insightful questions and suggestions, and we look forward to incorporating these improvements in the final version.

Official Review of Submission3583 by Reviewer pFMx

Official Review by Reviewer pFMx 20 Mar 2026, 09:08 (modified: 15 May 2026, 19:30)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer pFMx, Authors

Revisions (/revisions?id=R939GZNhSp)

Paper Summary:

This paper does not propose a new recommendation model. Instead, it provides a systematic analysis of the differences between generative recommendation and item-ID recommendation in terms of memorization and generalization. The authors first partition test instances based on item transition patterns to distinguish more memorization-oriented cases from more generalization-oriented ones. They then compare generative recommendation and traditional sequential recommendation models across these subsets on multiple datasets, and further offer a token-level explanation for why generative recommendation may perform better in certain generalization scenarios. Finally, the paper introduces a memorization-aware indicator to combine the two types of models and demonstrate their complementarity.

Paper Strengths:

1. The main contribution is analytical rather than model-centric, and I believe this is valuable for the community. The paper addresses a more fundamental question than simply which model gets a higher score: what generative recommendation is actually good at and where its limits are.
2. The analysis framework is fairly complete. The paper moves from item transition patterns to token-level explanation in a coherent way rather than presenting disconnected observations.
3. The experiments cover multiple public datasets, which gives the analysis broader support than many phenomenon-only papers.
4. The paper does not stop at describing differences. By adding a simple indicator to exploit the complementarity of the two model families, it also provides a practical takeaway.
5. Overall, the paper offers insights that are likely to remain useful, because it helps clarify the capability boundaries of generative recommendation.

Paper Weaknesses:

1. The memorization / generalization categorization depends on a rule system defined by the authors, so the conclusions are partly conditioned on this analytical framework.
2. Although the selected representative models are reasonable, the model coverage on both the generative recommendation side and the item-ID side is still somewhat limited.
3. The final indicator / ensemble part feels more like a proof of complementarity than a strong standalone methodological contribution.

4. The presentation has visible maturity issues, including leftover template text, which weakens the overall polish of the paper.
5. Some of the analytical claims would become even stronger with a few concrete case-based illustrations in addition to the aggregate statistics.

Resubmission:

According to the available submission information, this paper does not appear to be a resubmission, so this field is not applicable.

Questions And Suggestions For Rebuttal:

1. It would be helpful to clarify whether the main observations remain stable when replacing the current representative models with other generative recommendation or item-ID models.
2. Since the categorization framework is central to the paper, please state more directly the design rationale behind these rules and their possible limitations.
3. If the authors have already inspected representative cases, a brief example or two in the rebuttal would make the token-level explanation easier to appreciate.
4. I interpret the value of the indicator mainly as evidence of complementarity. It would help to state this more explicitly so that readers do not misread the paper as primarily an ensemble-method paper.

Relevance: 4: High - The work is relevant to the Research track of KDD and is of broad interest to the community

Novelty: 3: Moderate - The paper introduces a new and interesting idea or approach that adds value to the field. The contribution is original and represents an advancement of existing knowledge, demonstrating solid innovation and creativity.

Technical Quality: 4: High - The paper exhibits a high level of technical quality with a rigorous and well-executed methodology and analysis. The results are highly reliable, well-supported, and thorough. The work demonstrates technical excellence and sets a high standard for quality in the field.

Presentation: 2: Low - The paper has noticeable issues with clarity and coherence. The writing may contain several grammatical and typographical errors. Figures and tables are present but may not be well-integrated or effectively used. The presentation allows for understanding but requires effort from the reader.

Reproducibility: 3: Moderate - The paper provides a clear and detailed description of the methods, data, and procedures used. Supplementary materials, such as datasets and code, are available and sufficiently documented. Reproducing the results would be feasible with the provided information, though some effort may still be required.

Reviewer Confidence: 3: Moderate - The reviewer has a good understanding of the subject area and is familiar with the research methods and context of the paper. They feel confident in their ability to accurately assess the quality and significance of the work. Their evaluation is based on a solid grasp of the content and context.

Ethics Review Flag: No

Ethics Review Description: NA

**Rebuttal by Authors**

Rebuttal

by Authors ([👤 Yijie Ding \(/profile?id=~Yijie_Ding2\)](/profile?id=~Yijie_Ding2), [👤 Zitian Guo \(/profile?id=~Zitian_Guo1\)](/profile?id=~Zitian_Guo1), [👤 Jiacheng Li \(/profile?id=~Jiacheng_Li2\)](/profile?id=~Jiacheng_Li2), [👤 Letian Peng \(/profile?id=~Letian_Peng1\)](/profile?id=~Letian_Peng1), +7 more (/group/edit?id=KDD.org/2026/Research_Track_Cycle_2/Submission3583/Authors))

11 Apr 2026, 04:50 (modified: 11 Apr 2026, 04:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (</revisions?id=7ACAiv30aD>)

Rebuttal:**1. Do the observations hold to other model variants?**

Please see our response for reviewer xdQP.

2. Design rationale and limitations of the categorization framework

Intuitively, our design principle is to identify the most fundamental patterns a model can recall or compose from sequential dependencies. For example, as discussed in our response to Reviewer L7yC (Q2), 'symmetry' requires inferring swapped interaction order, which is crucial in real-world recommendation. The main advantages of our rule-based framework are its simplicity and scalability to large recommendation scenarios, since the categorization is very computationally efficient vs. alternatives such as counterfactual memorization. Moreover, our experiments show that the simple categorization effectively captures the consistent difference in GR and item ID-based models.

Nevertheless, there are also limitations, which we leave to future works:

- Our definition is based on the existence of a pattern, rather than the fine-grained occurrence frequency of memorization or generalization patterns.
- The rule-based definition does not establish a causal connection between a model's prediction and the corresponding training data

3. Qualitative examples of token-level explanations

Example from Beauty'14: S = SASRec, T = TIGER, N = NDCG@10

Item-level generalization reduces to token memorization

No.19159 Vitamin C Serum → Argan Oil Hair Treatment (S-N = 0.0, T-N = 1.0) This transition does not appear in training and requires multi-hop generalization (transitivity_3 + 2nd-symmetry_3). However, training contains 23 parallel transitions sharing the same 2-gram prefix pair [139, 310] → [159, 310] (e.g., VC Serum variants → Argan Oil variants). These allow TIGER to learn the pattern at the prefix level, enabling item-level generalization via token memorization

Token memorization dilutes item memorization

No.5873 Emu Oil Skin Repair → Organic Facial Moisturizer (S-N = 1.0, T-N = 0.0) This transition appears in all occurrences of the context item ($\phi = 1.0$). However, the dominant prefix transitions for [196, 310] → [234, 310] correspond to skin-repair → mask products, introducing 10 competing items ($\psi = 0.01$). This dilutes the probability for TIGER to generate the target item 'facial moisturizer'

4. The indicator mainly serves as evidence of complementarity

We thank the reviewer for this suggestion. We will make the motivation and contribution of the adaptive ensemble more explicit

Official Review of Submission3583 by Reviewer 6VL9

Official Review by Reviewer 6VL9  20 Mar 2026, 03:57 (modified: 15 May 2026, 19:30)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer 6VL9, Authors

 Revisions (/revisions?id=wa3E7uwdtC)

Paper Summary:

This paper provides a systematic comparison between GRs with SIDs and traditional recommendation models with atomic item IDs. It introduces a new rule to categorize interactions based on item transition patterns and demonstrates the benefits of GRs and traditional recommendation models with this type of data, separately.

Paper Strengths:

- Identifying the functional divergence between GR and traditional recommendation paradigms is crucial.
- The overall writing is a pleasure to read and easy to understand, with insightful figures and explanations.
- Extensive experiments across two backbone models and multiple real-world datasets demonstrate the effectiveness of this framework.

Paper Weaknesses:

- In my view, the core contribution of this work lies in the criteria for categorizing data instances and the subsequent findings regarding GRs versus traditional recommenders. However, the limited optimization of model architectures makes the paper more suited for the benchmark track rather than the research track.
- Experiments were conducted on small-scale datasets. To the best of my knowledge, the largest dataset used in the experiments contains fewer than 10 million interactions, which is significantly smaller than those typically used in industry (e.g., ~14 billion in [1]). The authors should provide additional evidence to demonstrate the scalability of the findings in this paper.
- As mentioned in Section 4.3.2 of this paper, while the shared structure of semantic prefixes provides a scaffolding for generalization, it also introduces interference from other items within the same semantic cluster when the model attempts to capture the uniqueness of a specific item. A key question remains: if the SID collisions are resolved using recent state-of-the-art strategies (e.g., [1, 2]), could GR models significantly close the gap in memorization performance?
- While SASRec and TIGER are highly representative of traditional and generative paradigms, respectively, the current selection of baselines is somewhat limited. Incorporating more backbones is essential to fully demonstrate the scalability of the findings.

[1] FORGE: Forming Semantic Identifiers for Generative Retrieval in Industrial Datasets

[2] Purely Semantic Indexing for LLM-based Generative Recommendation and Retrieval

Resubmission:

Not a Resubmission.

Questions And Suggestions For Rebuttal:

See the weakness.

Relevance: 4: High - The work is relevant to the Research track of KDD and is of broad interest to the community

Novelty: 3: Moderate - The paper introduces a new and interesting idea or approach that adds value to the field. The contribution is original and represents an advancement of existing knowledge, demonstrating solid innovation and creativity.

Technical Quality: 3: Moderate - The paper demonstrates solid technical quality with a sound methodology and thorough analysis. The results are reliable and well-supported. There may be minor issues, but they do not significantly undermine the overall quality. The work is competently executed and meets acceptable standards.

Presentation: 3: Moderate - The paper is organized and generally clear. The writing is mostly free of grammatical and typographical errors, making it easy to read. Figures and tables are effectively used to support the text. The presentation facilitates understanding and conveys the key points effectively.

Reproducibility: 3: Moderate - The paper provides a clear and detailed description of the methods, data, and procedures used. Supplementary materials, such as datasets and code, are available and sufficiently documented. Reproducing the results would be feasible with the provided information, though some effort may still be required.

Reviewer Confidence: 4: High - The reviewer is an expert in the subject area and has extensive knowledge of the research methods and context of the paper. They are highly confident in their ability to provide an accurate and thorough assessment. Their evaluation is based on deep expertise and a comprehensive understanding of the work.

Ethics Review Flag: No

Ethics Review Description: None



Rebuttal by Authors

Rebuttal

by Authors (Yijie Ding (/profile?id=~Yijie_Ding2), Zitian Guo (/profile?id=~Zitian_Guo1), Jiacheng Li (/profile?id=~Jiacheng_Li2), Letian Peng (/profile?id=~Letian_Peng1), +7 more (/group/edit?id=KDD.org/2026/Research_Track_Cycle_2/Submission3583/Authors))

11 Apr 2026, 04:51 (modified: 11 Apr 2026, 04:59)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=tX9Gv1M9H4)

Rebuttal:

1. Contributions to the Research Track

Rather than proposing new architectures simply to push state-of-the-art metrics, our paper aims to bridge a critical gap: the lack of analytical frameworks to understand why generative recommendation (GR) models demonstrate specific performance advantages over item ID-based models. To this end, we present a quantitative framework to expose these functional discrepancies and provide mechanistic evidence explaining GR's strong generalization capabilities. We believe these contributions would help guide future model designs in the research community.

2. How scalable are the findings to larger datasets?

We thank the reviewer for this important question regarding the applicability of our framework. To the best of our knowledge, we selected seven of the most commonly used public datasets in recent GR work from 2023–2025 for this study. Unfortunately, due to the amount of computational resources required for industry datasets, we would leave this as an exciting future work.

3. Would resolving SID collision significantly improve memorization?

We thank the reviewer for this insightful question and for pointing us to FORGE and PSID. We evaluated PSID [3] and report the results in our response to reviewer xdQP. We observe that resolving SID-level collisions does not directly lead to a significant improvement in memorization. We believe that it's because PSID's non-nearest centroid selection, or FORGE's KNN/random reassignment, mainly resolves leaf-node collisions, but can increase the codebook density at prefix layers. This increased prefix sharing raises the token memorization ratio and therefore strengthens the dilution effect on item memorization (Section 4.3).

To verify this, the token memorization ratio table of PSID (shown below) exhibits a very similar reduction pattern on the Beauty dataset to that of TIGER (Figure 4 in the paper), despite using one fewer token.

| token_cat | 3-gram | 2-gram | 1-gram | unseen |
|---------------|--------|--------|--------|--------|
| symmetry | 46.6 | 23.5 | 29.9 | 0.0 |
| transitivity | 20.2 | 23.9 | 55.9 | 0.0 |
| 2nd-symmetry | 17.5 | 22.8 | 59.7 | 0.0 |
| uncategorized | 0.0 | 11.5 | 88.5 | 0.0 |

[3] Purely Semantic Indexing for LLM-based Generative Recommendation and Retrieval

4. Do the findings generalize to other model backbones?

Please refer to our response for reviewer xdQP.



Official Comment by Reviewer 6VL9

Official Comment by Reviewer 6VL9 📅 16 Apr 2026, 05:11

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you for the detailed response. I have carefully reviewed the rebuttal and will increase my score accordingly.

Regarding the scalability of the proposed approach, it would be beneficial to further validate the method on larger public research datasets, such as MovieLens 10M or Amazon Books. In my experience, experiments on these datasets can be conducted with a single A100 GPU and would help strengthen the empirical evidence.

I am inclined to support the acceptance of this paper if the authors could incorporate the clarifications and improvements in the rebuttal into the revised version.



➔ *Replying to Official Comment by Reviewer 6VL9*

Official Comment by Authors

Official Comment

by Authors (👁️ [Yijie Ding \(/profile?id=~Yijie_Ding2\)](/profile?id=~Yijie_Ding2), [Zitian Guo \(/profile?id=~Zitian_Guo1\)](/profile?id=~Zitian_Guo1), [Jiacheng Li \(/profile?id=~Jiacheng_Li2\)](/profile?id=~Jiacheng_Li2), [Letian Peng \(/profile?id=~Letian_Peng1\)](/profile?id=~Letian_Peng1), +7 more (/group/edit?id=KDD.org/2026/Research_Track_Cycle_2/Submission3583/Authors))

📅 18 Apr 2026, 01:40

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We are glad that our responses have addressed your concerns, and we sincerely appreciate your suggestion to further validate the scalability of our findings.

Upon receiving your comment, we immediately began processing larger datasets (such as Amazon Books) and running evaluations. Unfortunately, due to resource constraints, we were unable to complete sufficient hyperparameter tuning to ensure a fair and reliable comparison

Hosting a Venue (/[venue](#))? The brief two-day timeframe. started/[frequently-asked-questions](#))

[id=OpenReview.net/Support](#)) we commit to completing these experiments with thorough tuning and including the large-scale

dataset results to further strengthen our findings. In the meantime, we will make sure to

incorporate the clarifications and improvements from the rebuttal into the revised manuscript.

[Sponsors \(/sponsors\)](#) [Terms of Use \(/legal/terms\)](#)

We thank you once again for your valuable feedback and your dedication to the rigor and

comprehensiveness of our experimental evaluation.

[\[Homepage\]\(/\)](#)

[OpenReview \(/about\)](#) is a long-term project to advance science through improved peer review with legal nonprofit status. We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2026 OpenReview