

[← Go to ACL ARR 2025 October homepage \(/group?id=aclweb.org/ACL/ARR/2025/October\)](#)

2 Versions ▾

# Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders



*Yupeng Hou (/profile?id=~Yupeng\_Hou1), Jiacheng Li (/profile?id=~Jiacheng\_Li2), Xiangjun Fu (/profile?id=~Xiangjun\_Fu1), Zhankui He (/profile?id=~Zhankui\_He1), An Yan (/profile?id=~An\_Yan1), Xiusi Chen (/profile?id=~Xiusi\_Chen1), Julian McAuley (/profile?id=~Julian\_McAuley1)*

07 Oct 2025 (modified: 17 Mar 2026) ACL ARR 2025 October Submission

October, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers

Revisions (/revisions?id=VEjHZ0vLQG) CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

## Abstract:

Feature engineering has long been central to recommender systems, yet effectively leveraging textual item features remains challenging. Recent advances in large language models (LLMs) have enabled their use as semantic encoders for recommendation, but their roles and behaviors in this setting are still not well understood. Prior studies often rely on general-purpose embedding benchmarks (e.g., MTEB) when selecting LLMs, overlooking the unique characteristics of recommendation tasks. To address this gap, we introduce BLAIR, a comprehensive benchmark for evaluating LLMs as semantic encoders in recommendation scenarios. We contribute (1) a new large-scale Amazon Reviews 2023 dataset with over 570 million reviews and 48 million items, (2) a unified benchmark covering sequential recommendation, collaborative filtering, and product search, and (3) a new complex-query product search task featuring both semi-synthetic and real-world evaluation datasets. Experiments with 11 leading LLMs show that their rankings on BLAIR show little correlation with MTEB, highlighting the unique challenges of semantic encoding in recommendation.

**Paper Type:** Long

**Research Area:** Resources and Evaluation

**Research Area Keywords:** benchmarking, retrieval, recommender system, sentence embedding

**Contribution Types:** Data resources, Data analysis

**Languages Studied:** English

**Reassignment Request Area Chair:** This is not a resubmission

**Reassignment Request Reviewers:** This is not a resubmission

**A1 Limitations Section:** This paper has a limitations section.

**A2 Potential Risks:** Yes

**A2 Elaboration:** We have a section named "Ethical Considerations"

**B Use Or Create Scientific Artifacts:** Yes

**B1 Cite Creators Of Artifacts:** Yes

**B1 Elaboration:** We cite used datasets

**B2 Discuss The License For Artifacts:** No

**B2 Elaboration:** We haven't, but all the used datasets are all widely used for research purpose.

**B3 Artifact Use Consistent With Intended Use:** N/A

**B4 Data Contains Personally Identifying Info Or Offensive Content:** Yes

**B4 Elaboration:** Ethical Considerations

**B5 Documentation Of Artifacts:** N/A

**B6 Statistics For Data:** Yes

**B6 Elaboration:** 3, 4

**C Computational Experiments:** Yes

**C1 Model Size And Budget:** Yes

**C1 Elaboration:** 5

**C2 Experimental Setup And Hyperparameters:** Yes

**C2 Elaboration:** Appendix

**C3 Descriptive Statistics:** Yes

**C3 Elaboration:** Appendix

**C4 Parameters For Packages:** N/A

**D Human Subjects Including Annotators:** No

**D1 Instructions Given To Participants:** N/A

**D2 Recruitment And Payment:** N/A

**D3 Data Consent:** N/A

**D4 Ethics Review Board Approval:** N/A

**D5 Characteristics Of Annotators:** N/A

**E Ai Assistants In Research Or Writing:** Yes

**E1 Information About Use Of Ai Assistants:** No

**E1 Elaboration:** We use ChatGPT for polishing the writing

**Author Submission Checklist:** yes

**Association For Computational Linguistics - Blind Submission License Agreement:**  On behalf of all authors, I agree

**TLDR:**  BLAIR is a comprehensive benchmark that evaluates LLMs as semantic encoders for recommendation, introducing a new large-scale Amazon Reviews 2023 dataset and a complex-query product search task to reveal that LLM performance in recommendation differs from general text embedding benchmarks.

**Preprint:**  no

**Preprint Status:**  There is a non-anonymous preprint (URL specified in the next question).

**Existing Preprints:**  <https://arxiv.org/pdf/2403.03952> (<https://arxiv.org/pdf/2403.03952>)

**Preferred Venue:**  ACL

**Consent To Share Data:**  yes

**Consent To Share Submission Details:**  On behalf of all authors, we agree to the terms above to share our submission details.

**Submission Number:** 3654

Filter by reply type...  Filter by author...  Search keywords... Sort: Newest First

Everyone Submission3654... Submission3654 Area... Submission3654 Authors 8 / 8 replies shown

Submission3654... Program Chairs Submission3654... Submission3654...

Submission3654... Submission3654...

### Meta Review of Submission3654 by Area Chair 3noG

Meta Review by Area Chair 3noG  04 Dec 2025, 12:08 (modified: 17 Mar 2026, 09:28)

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

 Revisions (/revisions?id=nrMftxH2pd)

#### Metareview:

This paper proposed BLAIR, a comprehensive benchmark for **evaluating LLMs as semantic encoder in recommendation scenarios**. They contributed (1) a new large-scale dataset, Amazon Reviews 2023, (2) a unified benchmark covering three core recommendation scenarios, and (3) a novel complex-query product search subtask,

designed to evaluate models' ability to handle long, ambiguous queries.

**Summary Of Reasons To Publish:**

There was a consensus among all reviewers that this paper should be accepted.

**Summary Of Suggested Revisions:**


Please add the contents suggested by the reviewers in your final version.


**Overall Assessment:** 4 = Conference: I think this paper could be accepted to an \*ACL conference.

**Reported Issues:**  No

**Publication Ethics Policy Compliance:** I did not use any generative AI tools for this review

**Official Review of Submission3654 by Reviewer DyTW**

Official Review by Reviewer DyTW  12 Nov 2025, 19:16 (modified: 17 Mar 2026, 09:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer DyTW, Authors, Commitment Readers

 Revisions (/revisions?id=Ppf0fwPvWe)

**Paper Summary:**

This paper introduces BLAIR, a benchmark for evaluating LLMs as semantic encoders in recommendation systems. The authors argue that existing work relies on general text embedding benchmarks (e.g., MTEB) for model selection, which may not reflect recommendation-specific requirements. They contribute: (1) a new Amazon Reviews 2023 dataset with 571M reviews and 48M items - substantially larger and more recent than previous versions, (2) a unified benchmark covering sequential recommendation, collaborative filtering, and product search across 14 datasets, and (3) a novel complex-query product search task with both semi-synthetic (Amazon-C4) and real-world (Reddit-Movie) evaluation sets. Experiments with 11 leading LLMs reveal that BLAIR rankings show little correlation with MTEB ( $\rho = -0.476$ ), scaling benefits diminish with downstream task complexity, and the semi-synthetic dataset correlates strongly ( $r = 0.94$ ) with real-world complex queries.

**Adequacy Of Revisions:**

N/A

**Summary Of Strengths:**

S1: Excellent clarity and presentation. The paper is very well written with clear motivation, problem formulation, and experimental setup. The flow from problem identification (general embedding benchmarks may not suit recommendation) to solution (domain-specific benchmark) to validation is easy to follow without requiring significant effort from readers.

S2: Significant and timely dataset contribution. The Amazon Reviews 2023 dataset addresses a real community need. It is substantially larger (3.18x more items, 2.58x more text tokens), extends coverage by 5 years (through Sep 2023), provides cleaner metadata with richer structured fields, and offers millisecond-precision timestamps versus day-level granularity. This will benefit recommendation research and broader NLP work on user-generated content and e-commerce.

S3: Well-designed benchmark addressing a genuine gap. The paper makes a compelling case that LLMs-as-semantic-encoders for recommendation require specialized evaluation distinct from general text embedding tasks. The comprehensive evaluation across 11 diverse LLMs and 14 datasets with three recommendation scenarios demonstrates thoroughness. The use of adapter layers to control downstream model parameters ensures fair comparison, isolating semantic encoder quality from model capacity.

S4: Novel complex-query product search task with strong validation. Introducing evaluation for long, descriptive, ambiguous queries addresses emerging real-world needs (conversational shopping assistants). The dual evaluation approach - semi-synthetic Amazon-C4 and real-world Reddit-Movie - is methodologically sound, and their strong correlation ( $r = 0.94$ ) validates using synthetic data as a cost-effective evaluation proxy.

S5: Important empirical findings with actionable insights. The finding that BLAIR and MTEB rankings show little correlation ( $\rho = -0.476$ ) is significant for practitioners currently relying on MTEB for model selection. The observation about scaling behavior and text-embedding-3-large's strong generalization raise interesting questions about

benchmark-specific optimization versus true generalization. The commitment to release dataset, toolkit, and scripts enhances reproducibility and community benefit.

#### Summary Of Weaknesses:

W1: Missing explanation for counter-intuitive scaling behavior. Lines 439-457 report that larger semantic encoders show diminishing returns for sequential recommendation (Transformer decoder downstream) compared to collaborative filtering (linear layer downstream). This is counter-intuitive and significant, yet no hypothesis or reasoning is provided. Why would larger encoders help less with complex downstream models? This finding has implications for architecture design and deserves at least a hypothesis about the underlying mechanism.

W2: Lack of hypothesis for MTEB-BLAIR mismatch. The paper's key finding (lines 418-437) is that rankings on MTEB and BLAIR show little correlation ( $\rho = -0.476$ ,  $p = 0.233$ ). Specifically, text-embedding-3-large ranks 42nd on MTEB but 1st (Borda) on BLAIR. While the paper mentions possible factors (clustering/summarization tasks, discriminative representations, overfitting), these remain speculative. What hypothesis explains this mismatch? Is it about task characteristics, training objectives, or something else? A clearer analysis of why these benchmarks diverge would strengthen the contribution.

W3: Questionable claim about description redundancy. Lines 509-518 conclude that item descriptions don't consistently improve performance, attributing this to descriptions being "redundant" because "much of the factual content may already be captured by the world knowledge encoded in LLMs." I disagree with this reasoning. A title like "cotton black t-shirt" is generic, but descriptions specifying "100% cotton" vs "75% cotton blend" contain crucial product-specific attributes that directly impact user preference and are NOT general world knowledge. The experiment only tests 3 models on 4 datasets - insufficient to support such a broad conclusion. This needs either more comprehensive experiments or acknowledgment as an open problem.

#### Comments Suggestions And Typos:

N/A

**Confidence:** 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

**Soundness:** 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.

**Excitement:** 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.

**Overall Assessment:** 4 = Conference: I think this paper could be accepted to an \*ACL conference.

#### Ethical Concerns:

None.

The authors have included a thoughtful Ethical Considerations section (lines 593-619) addressing data privacy and potential societal impacts. They clarify that the Amazon Reviews 2023 dataset contains only publicly available information that users explicitly chose to share, excludes user metadata to prevent profiling, and focuses on item-level data. All benchmark tasks use publicly released datasets following original usage policies. The authors appropriately acknowledge that recommender systems can reinforce biases or create filter bubbles, and note their benchmark evaluates semantic encoding capabilities rather than deployment-ready systems. They encourage researchers to consider fairness and transparency when developing real-world applications.

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

**Software:** 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

**Publication Ethics Policy Compliance:** I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



## Thanks for your review!

Official Comment

by Authors ( Jiacheng Li (/profile?id=~Jiacheng\_Li2), An Yan (/profile?id=~An\_Yan1), Zhankui He (/profile?id=~Zhankui\_He1), Yupeng Hou (/profile?id=~Yupeng\_Hou1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3654/Authors))

23 Nov 2025, 06:30 (modified: 17 Mar 2026, 09:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer DyTW, Commitment Readers

Revisions (/revisions?id=2qpHTYohM9)

### Comment:

We thank the reviewer for the time and effort spent in reviewing our paper. We made several discussions on the reviewer's questions below.

#### W1: Discussion on why scaling semantic encoders yields limited gains on complex tasks.

Thank you for raising such an insightful discussion. Our current hypotheses are as follows:

- **Differences in the nature of the tasks:** The representational capabilities required differ between tasks.
  - **Collaborative filtering** is primarily about encoding similarity. A stronger semantic encoder can represent item similarities in a richer and more accurate manner, which directly translates into performance gains in CF as the encoder scales up. For example, Sheng et al. [1] found that stronger semantic encoders yield representations whose distribution is more closely aligned with those learned from pure interaction data.
  - **Sequential recommendation** is essentially a sequence modeling problem that requires memorizing and reproducing behavior patterns. To achieve this, item representations may require strong **discriminative capability** (akin to functioning as unique indices). Representations produced by even a smaller semantic encoder might already be sufficiently distinguishable for a complex downstream Transformer model. Consequently, scaling the semantic encoder does not significantly enhance discriminability from the perspective of the downstream model, leading to smaller gains compared to CF. A similar observation was reported by Hou et al. [2], who showed that treatments such as whitening enhance the discriminative power of item representations and improve sequential recommendation performance.
- **Sequentially dependent neural network systems and scaling behaviors:**
  - Our setup functions as a two-stage system: a first-stage semantic encoder whose output is consumed by a second-stage model (a linear layer for CF, or a Transformer decoder for SeqRec).
  - We hypothesize that in such multi-stage systems, scaling the later-stage modules may diminish the impact of scaling earlier-stage modules. While there is limited systematic work on scaling laws in sequentially dependent systems, our findings may suggest that *where* we allocate additional parameters may matter as much as *how many* parameters we add.

We will incorporate this discussion into the revised paper and plan to investigate this phenomenon further in future work. Thanks again for the suggestion!

[1] Sheng et al. Language representations can be what recommenders need: Findings and potentials. ICLR 2025.

[2] Hou et al. Towards Universal Sequence Representation Learning for Recommender Systems. KDD 2022.

**W2: Discussion on the mismatch between BLAIR and MTEB benchmarks.**

Our main hypothesis is that some models may be overfitted to MTEB-style tasks, for example by increasing the proportion of similar tasks in their training data. There may be unintentional data leakage that influences MTEB performance. This could lead to strong performance on MTEB while failing to generalize to BLAIR, which emphasizes different retrieval and recommendation scenarios.

That said, we emphasize that these explanations are no more than guesses. Because most providers do not release training data, and some models do not even release their model parameters, it is difficult to draw firm conclusions about overfitting or contamination. In the paper, we therefore present the MTEB-BLAIR mismatch primarily as an empirical finding.

**W3: Discussion on the claim of the item metadata experiments.**

Thank you for offering this valuable perspective on our findings! We acknowledge that our original explanation reflects only one possible interpretation of the results, and your comments help us make the discussion more comprehensive. We will revise the manuscript to add this interpretation as well, and clearly state that feature engineering for semantic encoders remains an open problem. We also hope that the provided BLAIR benchmark can serve as a useful resource for researchers interested in further studying this question.

**Official Review of Submission3654 by Reviewer ho3j**

Official Review by Reviewer ho3j 📅 11 Nov 2025, 04:08 (modified: 17 Mar 2026, 09:28)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer ho3j, Authors, Commitment Readers

📄 Revisions (/revisions?id=FwiKSdLd5r)

**Paper Summary:**

This paper presents BLAIR (Bridging Language and Items for Retrieval and Recommendation), a comprehensive benchmark for evaluating Large Language Models (LLMs) as semantic encoders in recommendation scenarios.

The authors argue that conventional embedding benchmarks such as MTEB fail to capture the unique requirements of recommendation tasks, where text representations are used as model inputs (rather than final outputs) and item descriptions tend to be short and noisy.

To address this, BLAIR contributes:

1. A new large-scale dataset, Amazon Reviews 2023, containing 570M reviews and 48M items, with cleaned metadata.
2. A unified benchmark covering three core recommendation scenarios—sequential recommendation, collaborative filtering, and product search—along with 11 strong baselines.
3. A novel complex-query product search subtask, constructed from both semi-synthetic Amazon-C4 data and real-world forum queries, designed to evaluate models' ability to handle long, ambiguous queries.

Experiments with 11 leading LLMs show that model rankings on BLAIR are largely uncorrelated with MTEB, underscoring the domain-specific challenges of using LLMs as semantic encoders in recommendation.

**Summary Of Strengths:**

1. A high-quality and timely benchmark that fills an important gap between text embedding research and recommender systems.
2. The paper provides three well-defined evaluation scenarios with reproducible baselines, offering systematic coverage of practical recommendation use cases.
3. The Amazon Reviews 2023 dataset is a valuable resource for the research community, addressing long-standing issues of outdated or noisy metadata in prior Amazon corpora.
4. The complex query product search task is thoughtfully designed and relevant to emerging retrieval applications. It can be considered as a new mainstream approach to synthesis queries.

5. Results are well-analyzed, clearly demonstrating that semantic encoding for recommendation differs fundamentally from general-purpose embedding tasks.
6. The paper is clearly structured, well-written, and easy to follow.
7. They open-source their resources.

#### Summary Of Weaknesses:

The research question is al dante. The benchmark is comprehensive, methodological novelty is limited—the contribution is mainly infrastructural rather than algorithmic.

The scaling-law observation mentioned in the conclusion is interesting but somewhat underexplored; it deserves more systematic analysis.

The relation to pretraining objectives (e.g., how SFT or RL-tuned LLMs differ) could be discussed to provide more insight.

Overall, I do not see very significant weakness from this paper.

#### Comments Suggestions And Typos:

1. Consider adding error analysis or qualitative examples showing failure cases of LLM encoders on BLAIR tasks.
2. Authors can consider expanding on the insight behind the weak MTEB correlation—which task characteristics drive this divergence? how BLAIR can inform future model design? (e.g., for domain-adaptive semantic encoders).

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

**Soundness:** 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.

**Excitement:** 4.5

**Overall Assessment:** 4 = Conference: I think this paper could be accepted to an \*ACL conference.

#### Best Paper Justification:

If the paper is submitted to a RecSys-focused conference, I would recommend considering it in the award tier.

#### Ethical Concerns:

There are no concerns with this submission

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** Yes

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** Somehow

**Knowledge Of Authors Guess:** Consider the scale of the benchmark and experiments, I would guess if it is from Julian McAuley's Lab. But this will not affect my justification.


**Reviewer Certification:** I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

**Publication Ethics Policy Compliance:** I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



### Thanks for your review! (Part 1/2)

Official Comment

by Authors ( Jiacheng Li (/profile?id=~Jiacheng\_Li2), An Yan (/profile?id=~An\_Yan1), Zhankui He (/profile?id=~Zhankui\_He1), Yupeng Hou (/profile?id=~Yupeng\_Hou1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3654/Authors))

📅 23 Nov 2025, 06:31 (modified: 17 Mar 2026, 09:28)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ho3j, Commitment Readers

📄 Revisions (/revisions?id=rzyBdVv3xF)

### Comment:

Thank you for the positive feedback in our work! Below, we discuss several points inspired by the reviewer's comments.

#### (1) The benefits of scaling semantic encoders diminish as the downstream task becomes more complex.

We thank the reviewer for highlighting this interesting observation. We have a few hypotheses that may help explain this phenomenon:

- **Differences in the nature of the tasks:** The representational capabilities required differ between tasks.
  - **Collaborative filtering** is primarily about encoding similarity. A stronger semantic encoder can represent item similarities in a richer and more accurate manner, which directly translates into performance gains in CF as the encoder scales up. For example, Sheng et al. [1] found that stronger semantic encoders yield representations whose distribution is more closely aligned with those learned from pure interaction data.
  - **Sequential recommendation** is essentially a sequence modeling problem that requires memorizing and reproducing behavior patterns. To achieve this, item representations may require strong **discriminative capability** (akin to functioning as unique indices). Representations produced by even a smaller semantic encoder might already be sufficiently distinguishable for a complex downstream Transformer model. Consequently, scaling the semantic encoder does not significantly enhance discriminability from the perspective of the downstream model, leading to smaller gains compared to CF. A similar observation was reported by Hou et al. [2], who showed that treatments such as whitening enhance the discriminative power of item representations and improve sequential recommendation performance.
- **Sequentially dependent neural network systems and scaling behaviors:**
  - Our setup functions as a two-stage system: a first-stage semantic encoder whose output is consumed by a second-stage model (a linear layer for CF, or a Transformer decoder for SeqRec).
  - We hypothesize that in such multi-stage systems, scaling the later-stage modules may diminish the impact of scaling earlier-stage modules. While there is limited systematic work on scaling laws in sequentially dependent systems, our findings may suggest that *where* we allocate additional parameters may matter as much as *how many* parameters we add.

We will incorporate this discussion into the revised paper and plan to investigate this phenomenon further in future work.

[1] Sheng et al. Language representations can be what recommenders need: Findings and potentials. ICLR 2025.

[2] Hou et al. Towards Universal Sequence Representation Learning for Recommender Systems. KDD 2022.

#### (2) Discussion on the relation to pretraining objectives (e.g., how SFT or RL-tuned LLMs differ).

This point is quite pioneering. To the best of our knowledge, although SFT has been widely adopted for tuning embedding models, the potential of RL for training embedding models remains largely underexplored. We believe this is an interesting direction for future work, not only in the context of our benchmark but for embedding model research more broadly.



**Thanks for your review! (Part 2/2)**

## Official Comment

by Authors (👁️ Jiacheng Li (/profile?id=~Jiacheng\_Li2), An Yan (/profile?id=~An\_Yan1), Zhankui He (/profile?id=~Zhankui\_He1), Yupeng Hou (/profile?id=~Yupeng\_Hou1), +3 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3654/Authors))

📅 23 Nov 2025, 06:31 (modified: 17 Mar 2026, 09:28)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ho3j, Commitment Readers

📄 Revisions (/revisions?id=WLQCd9JH8w)

**Comment:****(3) Error analysis or qualitative examples showing failure cases.**

Thanks for the great suggestion! We provide an example below that demonstrates a failure case of current semantic encoders on the complex-query product-search task in the Reddit Movie dataset.

The user query is: *"Over the Top Bonkers Action Movies?. I recently saw the new Suicide Squad movie and loved it. It had all the qualities I like in an action movie. Violence, shock factor, comedy and overall bizarreness. What are some other action movies with these qualities? Deadpool, Kickass, Turbo Kid, Tropic Thunder and Drive Angry are some of my other favorites."*

The top-recommended movies (ground truth) include: *Nobody (2021), Hardcore Henry (2015), Slow Moe (2010), and Dredd (2012).*

Using GritLM-7B, which is the best-performing semantic encoder on the Reddit Movie dataset, we show its top-3 retrieved movies: *Ninjas vs. Zombies (2008), Cowboys vs. Zombies (2014), and Cowboys vs. Vampires (2010).*

We can see that these retrieved movies share some superficial characteristics with the query: (1) many contain "vs.", which may correlate with the phrase "action movie" in the query; (2) they feature fantastical characters such as zombies and ninjas, which may loosely correspond to the idea of "bizarreness" in the query.

This case highlights several observations:

1. There remains substantial room for improvement in understanding complex user queries, given that even the best semantic encoder achieves only  $NDCG@100 = 0.0734$  over 50k candidate movies.
2. Popularity remains an important ranking signal (e.g., *Nobody* and *Hardcore Henry* are far more popular than the retrieved movies) and should be incorporated for complex-query product search.
3. More advanced techniques beyond surface-level semantic matching may be needed, such as explicit reasoning or deeper modeling of intent.

We will add more such qualitative examples and analyses in the next version. Thanks again for the suggestion!


**(4) Discussion on the mismatch between BLAIR and MTEB benchmarks.**

Our hypothesis is that building a strong semantic encoder requires three key capabilities:

1. Accurately modeling semantic similarity, which is essential for tasks such as collaborative filtering and short-query product search.
2. Producing distinguishable and informative feature representations, which supports tasks like sequential recommendation.
3. Exhibiting strong instruction-following and reasoning abilities, which are crucial for handling complex-query product search.

**Official Review of Submission3654 by Reviewer JhKw**

Official Review by Reviewer JhKw  10 Nov 2025, 22:17 (modified: 17 Mar 2026, 09:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer JhKw, Authors, Commitment Readers

 Revisions (/revisions?id=rzFEBG1k8)

### Paper Summary:

The paper proposed BLAIR, a comprehensive benchmark built on a new large-scale Amazon Reviews 2023 dataset to evaluate large language models as semantic encoders for recommendation tasks including sequential recommendation, collaborative filtering, and short- and complex-query product search.

### Adequacy Of Revisions:

N/A

### Summary Of Strengths:

1. good motivation to benchmark LLM embedding in recommendation tasks using some relatively "fresh data"
2. interesting findings that general embedding benchmark does transfer to recommendation
3. important open-source data contribution to the community

### Summary Of Weaknesses:

1. Would be interesting to see the tradeoff between model performance and latency
2. The data only source from amazon review 2023, which may limit the generalization to other platform and recommendation scenario
3. complex-query product search task partly relies on semi-synthetic queries, so the gap between these queries and truly organic user queries may affect the reliability of evaluation
4. would be nice to include some traditional recommendation method performance, it would be interesting to see how they are compared to the semantic embeddings.

### Comments Suggestions And Typos:

N/A

**Confidence:** 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

**Soundness:** 4.5

**Excitement:** 4 = Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.

**Overall Assessment:** 4 = Conference: I think this paper could be accepted to an \*ACL conference.

### Ethical Concerns:

There are no concerns with this submission

**Needs Ethics Review:** No

**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.


**Publication Ethics Policy Compliance:** I did not use any generative AI tools for this review




**Thanks for your review!**

## Official Comment

by Authors ([👤](#) Jiacheng Li (/profile?id=~Jiacheng\_Li2), [An Yan \(/profile?id=~An\\_Yan1\)](#), [Zhankui He \(/profile?id=~Zhankui\\_He1\)](#), [Yupeng Hou \(/profile?id=~Yupeng\\_Hou1\)](#), [+3 more \(/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3654/Authors\)](#))

 23 Nov 2025, 06:32 (modified: 17 Mar 2026, 09:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer JhKw, Commitment Readers

 Revisions (/revisions?id=WyFI8ij4mo)

**Comment:**

We would like to thank the reviewer for highlighting the open-sourced resources and finding our findings interesting. Below we address the concerns carefully.

**W1: Tradeoff between model performance and latency.**

Thanks for raising this important point about latency. We mainly consider two scenarios:

1. Recommendation tasks such as collaborative filtering and sequential recommendation. In the BLAIR benchmark, we explicitly project the raw representations produced by the semantic encoders into a fixed dimension. This makes latency across different encoders largely comparable, since the item/user representations can be pre-computed and cached. Because latency is less of a concern in these settings, the benchmark results provide a reasonable estimate of the encoders' capabilities.
2. Retrieval tasks such as short- and complex-query product search. Here, latency primarily comes from encoding the user query, since item representations can also be pre-computed and cached. From the benchmark results, we observe that semantic encoders around 7B parameters generally perform best. Although some closed-source models do not disclose their parameter counts, using a 7B-scale model is often practical in real-world systems due to the substantial performance gains and the continually decreasing cost of serving LLMs in recent years.

**W2: Concerns on limiting generalization to other platforms (besides Amazon) and other recommendation scenario.**

We understand the concern regarding generalization, which is essential to ensure that a benchmark is not overly tailored to specific platforms or scenarios. To address this, we include several widely used public datasets beyond Amazon, such as MovieLens, Yelp, Book-Crossing, and Reddit. These datasets not only represent different platforms but also span diverse recommendation domains, including movies, restaurants, and books. We believe this diversity is helpful in improving the generalizability of the proposed BLAIR benchmark.

**W3: The gap between semi-synthetic queries and truly organic user queries in the complex-query product search task may affect the reliability of evaluation.**

Due to the lack of publicly available datasets for complex-query product search, we use semi-synthetic queries as a proxy. To assess the reliability of this evaluation setup, we examine the correlation of model rankings between the semi-synthetic Amazon-C4 dataset and a real-world complex-query dataset, Reddit-Movie.

As reported in lines 458–473, the NDCG@100 scores of all evaluated models on these two datasets exhibit a strong positive linear correlation of 0.94 ( $p < 0.01$ ). This indicates that Amazon-C4, despite being semi-synthetic and covering different domains from real-world queries, effectively captures key characteristics of complex-query tasks.

That said, we acknowledge that understanding and validating datasets for complex-query product search is still in its early stages. We hope that the BLAIR benchmark can serve as a useful resource and reference point for future research in this direction.

**W4: Results of traditional recommendation method performance.**

Thanks for the suggestion! We report performance of two widely studied sequential recommendation models GRU4Rec [1] and SASRec [2] as references.

Model (NDCG@10)	All_Beauty	Video_Games	Baby_Products
GRU4Rec	0.0022	0.0119	0.0067
SASRec	0.0065	0.0124	<b>0.0083</b>
UniSRec (SimCSE)	0.0232	0.0114	0.0069
UniSRec (text-emb-3-large)	<b>0.0237</b>	<b>0.0135</b>	0.0078

We observe that on smaller datasets (analogous to cold-start scenarios), semantic-encoder-based methods perform significantly better, whereas on larger datasets, traditional models such as SASRec tend to achieve stronger results.

It is important to note that, as a benchmark for selecting semantic encoders, achieving the absolute best downstream performance is not our primary goal. In practice, a selected semantic encoder is typically used to generate feature representations, which are then combined with traditional recommendation models (e.g., a mixture of learnable ID embeddings and semantic representations). Moreover, on larger datasets, traditional models often scale substantially in the number of learnable parameters, while semantic-encoder-based methods maintain a fixed parameter size. This makes direct comparisons less fair and should be interpreted with caution.

Therefore, we mainly treat the results of traditional models as reference points rather than the central focus of the benchmark. Nonetheless, we appreciate the reviewer's suggestion which has helped make our paper more complete.

[1] Hidasi et al. Session-based recommendations with recurrent neural networks. ICLR 2016.

[2] Kang and McAuley. Self-attentive sequential recommendation. ICDM 2018.

[About OpenReview \(/about\)](#)
[FAQ \(https://docs.openreview.net/getting-started/frequently-asked-questions\)](#)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](#)
[Contact \(/contact\)](#)

[Avenue \(Venues\)](#)
[Donate \(/donate\)](#)

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

[News \(/group?id=OpenReview.net/News&referrer=\)](#)
[Privacy Policy \(/legal/privacy\)](#)

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

[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