Language Models as Semantic Augmenters for Sequential Recommenders

Mahsa Valizadeh*, **Xiangjue Dong, Rui Tuo, James Caverlee**
Texas A&M University
{mvalizadeh, xj.dong, ruituo, caverlee}@tamu.edu

Abstract

Large Language Models (LLMs) excel at capturing latent semantics and contextual relationships across diverse modalities. However, in modeling user behavior from sequential interaction data, performance often suffers when such semantic context is limited or absent. We introduce **LaMAR**, a LLM-driven semantic enrichment framework designed to enrich such sequences automatically. LaMAR leverages LLMs in a few-shot setting to generate auxiliary contextual signals by inferring latent semantic aspects of a user's intent and item relationships from existing metadata. These generated signals, such as inferred usage scenarios, item intents, or thematic summaries, augment the original sequences with greater contextual depth. We demonstrate the utility of this generated resource by integrating it into benchmark sequential modeling tasks, where it consistently improves performance. Further analysis shows that LLM-generated signals exhibit high semantic novelty and diversity, enhancing the representational capacity of the downstream models. This work represents a new data-centric paradigm where LLMs serve as intelligent context generators, contributing a new method for the semi-automatic creation of training data and language resources.

Keywords: Large Language Models (LLMs), Semantic Augmentation, Sequential Behavior Modeling

1. Introduction

Capturing users' dynamic preferences through their historical behaviors has been a central focus of sequential recommender systems (Chen et al., 2018; Xie et al., 2022; Liu et al., 2024a; Lin et al., 2024). A persistent limitation is that these interaction histories are often represented as sparse sequences of item identifiers with minimal contextual information, making it difficult for models to infer nuanced user preferences, especially in cold-start scenarios or long-tail items. Meanwhile, large language models (LLMs) like ChatGPT and Gemini have demonstrated remarkable capabilities in capturing semantic relationships and generating coherent text based on minimal input (Achiam et al., 2023). Their ability to reason about real-world concepts and infer latent intent makes them promising tools for enhancing recommendation systems with richer semantic context.

In this work, we propose **LaMAR** (Language Model-Augmented Recommendation), which leverages the reasoning capabilities of LLMs to enrich item representations through automatically generated semantic features (see Figure 1). Instead of traditional feature engineering, LaMAR uses an LLM to generate auxiliary descriptions and latent signals that provide context to a user's interaction history. Concretely, given a user's past sequence of items, we prompt a large language model (with a carefully designed prompt and a few illustrative examples) to produce additional information, such as categorical tags or topics that summarize the user's interests, or an explanation of the possible intent or scenario underlying the sequence. These generated signals are then incorporated into the training of the sequential recommender (e.g., by encoding them into the model's input representations). By enriching each user's history with LLM-inferred context, the model gains a deeper understanding of preferences that would not be evident from the raw data alone. For instance, based on a user's last five product interactions, the LLM may infer an underlying intent (e.g., "planning a camping trip") or suggest a latent thematic category (e.g., "outdoor adventure gear").

We conduct extensive experiments to evaluate the effectiveness of our framework on multiple public sequential recommendation datasets, comparing against state-of-the-art sequential recommendation baselines. Our results show that adding LLM-generated signals to user-item sequences leads to consistent performance gains across all key ranking metrics. To understand the contributions of the generated signals, we perform a thorough analysis of their diversity and uniqueness, and demonstrate their semantic richness. We find that the LLM is capable of contributing a wide range of complementary information rather than simply echoing a user's existing history, which expands the knowledge base of the recommender. Furthermore, we explore fine-tuning the LLM using the augmented sequences and observe additional improvements. Our code and data are available here: https://github.com/mahsavalizadeh/LaMAR/.

In summary, our contributions are as follows:

* Corresponding author: mvalizadeh@tamu.edu

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10467–10484
11–16 May 2026.

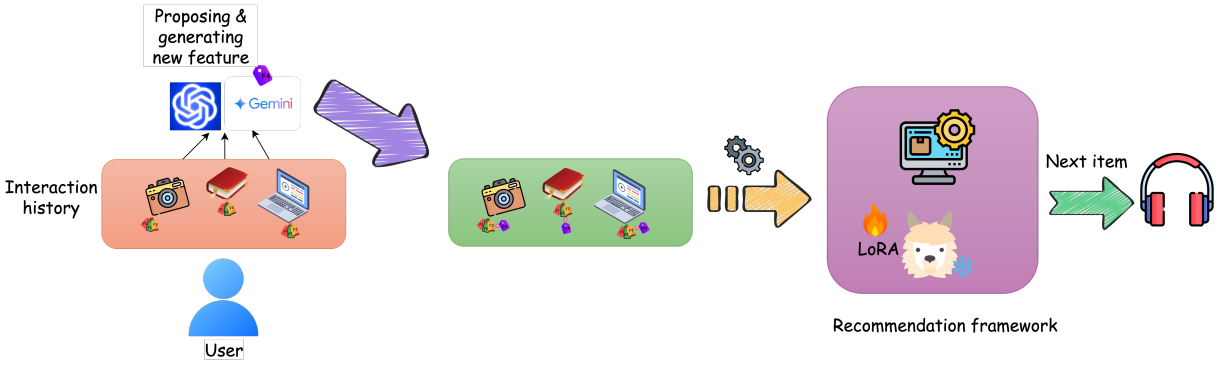


Figure 1: **Overview of the LaMAR framework.** It consists of two stages: (1) semantic signal generation using a prompted large language model, and (2) integration of the generated signals into a sequential recommendation model via full fine-tuning, or into the LLM through parameter-efficient LoRA tuning.

- We introduce a semantic augmentation framework that uses LLMs to generate informative and semantically diverse contextual signals for sequential recommendation. These signals are designed to capture semantic relationships and enhance the richness of the dataset, providing additional context that sequential models can leverage.
- We conduct a deeper analysis to assess the extent of feature redundancy and representation diversity. Specifically, we investigate how similar or distinct the generated features are, identifying potential overlaps and unique contributions.
- We show that incorporating these signals into existing sequential recommendation models yields significant improvements across multiple datasets.
- We explore the impact of data augmentation on sequential recommendation tasks while fine-tuning LLMs on both the enriched and original datasets.

2. Related Work

LLMs have recently transformed how we understand and represent semantics across a wide range of AI tasks, including recommendation. Their capacity for language understanding, content generation, and contextual reasoning has inspired many efforts to incorporate LLMs into recommender systems.

Early studies primarily leveraged LLMs for semantic augmentation, enriching item or user representations and generating additional training data in recommender systems (Lin et al., 2023; Liu et al., 2024b). For example, Brinkmann et al. (2023) use ChatGPT to extract attribute–value pairs from product descriptions, while KAR (Xi et al., 2024) and LLM-Rec (Lyu et al., 2023) leverage LLMs to produce auxiliary knowledge or enriched item descriptions. Other work explores broader content-enrichment strategies, including content summarization (Liu et al., 2024c), category descriptions (Yada and Yamana, 2024), taxonomy-guided augmentation (Liang et al., 2024), factual property extraction (Luo et al., 2024), and description-based augmentation for ID-based recommenders (Ren et al., 2024). Beyond feature enrichment, several approaches employ LLMs to produce synthetic interactions or simulate user behaviors (Wang et al., 2024a; Huang et al., 2025), construct structured semantic representations like knowledge graphs (Wang et al., 2024b; Liu et al., 2025) or provide reasoning paths (Bismay et al., 2025).

Before the emergence of LLM-based methods, architectures such as SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019) demonstrated strong performance through attention-based modeling of user-item interaction sequences. These and other methods – e.g., (Chen et al., 2018; Xie et al., 2022) – have typically modeled items simply by an ID, ignoring the rich semantic information contained in item content (e.g., text, images, or videos). Recent works have thus sought to bridge this gap by infusing item representations into sequential recommenders. Some explore *semantic IDs* (Singh et al., 2024), replacing simple item IDs with semantically-meaningful embeddings. Others leverage LLMs to directly model item content, e.g., RecFormer (Li et al., 2023) textualizes an item by flattening item characteristics (like title and description) into a language-based representation, while Rella (Lin et al., 2024) applies zero-/few-shot LLM reasoning for recommendation.

In contrast, our work departs from these constrained enrichment methods. Rather than limiting the LLM to predefined augmentation tasks, we ask it to identify and generate unconstrained seman-

Dataset	Category	Description
Office Products	Title	Ballet 2012 Square 12X12 Wall Calendar (Multilingual Edition)
	Brand	Browntrout Publishers (COR)
	Category	Office Products, Office & School Supplies, Calendars, Planners & Personal Organizers
	Product Function	Stylishly showcases beautiful ballet imagery while helping to organize dates and events throughout the year with its easy-to-read monthly grids.
Musical Instruments	Title	DW Drums 7000 Series Double Bass Drum Pedal
	Brand	Drum Workshop, Inc.
	Category	Musical Instruments Instrument Accessories Drum & Percussion Accessories Drum Set Accessories Drum Accessories Bass Drum Pedals
	Music Genre	Ideal for rock, metal, and jazz drumming styles, this pedal offers precision and responsiveness for both live performances and studio recordings, enhancing rhythmic expression.
Video Games	Title	America's Greatest Solitaire Games
	Brand	Wizardware Group
	Category	Video Games Mac Games
	Gameplay Experience Focus	Engaging solitaire gameplay featuring a variety of classic card games, strategic challenges, and customizable settings, perfect for casual gamers on Mac.
Industrial and Scientific	Title	Arctic Silver 5 AS5-3.5G Thermal Paste
	Brand	Arctic Silver
	Category	Industrial & Scientific Industrial Electrical Thermal Management Products Computer Heatsinks
	Primary Use Case	Thermal paste designed for computer heatsinks to improve heat transfer and cooling efficiency for optimal performance and longevity of electronic devices.
Arts, Crafts and Sewing	Title	Electric Laser Guided Scissors Stainless Steel Blades
	Brand	Salco
	Category	Arts, Crafts & Sewing Crafting Craft Supplies Cutting Tools Scissors
	Primary Craft Purpose	These laser-guided scissors provide precision cutting for various materials, enhancing crafting projects with clean and accurate results. Perfect for detailed and intricate designs.
Pet Supplies	Title	Hikari Usa Inc AHK01389 Staple 22lb, Medium
	Brand	Hikari Usa Inc.
	Category	Pet Supplies Fish & Aquatic Pets Food
	Pet's Specific Need	A balanced and nutritious staple food to support the health and growth of medium-sized fish in aquatic environments.

Table 1: Examples of LLM-generated semantic augmentations (highlighted in blue), including category and its specific value, for subsets of the Amazon Review Data (Ni et al., 2019): Office Products, Musical Instruments, Video Games, Industrial and Scientific, Arts, Crafts and Sewing, and Pet Supplies.

tic signals—from nuanced gameplay experiences to product-specific usage contexts—that may not be present in the original metadata. Further, these signals are integrated directly into the recommendation model in a flexible, end-to-end fashion, instead of being used merely as auxiliary text features. This design enables the recommender to dynamically leverage diverse, new semantic cues, offering a complementary and more adaptive framework for improving recommendation quality, especially for items with limited interactions.

3. Methodology of LaMAR

Our proposed framework, LaMAR (Language Model-Augmented Recommendation), consists of two primary stages: (1) a semantic signal generation pipeline that uses LLMs to enrich user-item sequences with contextual cues, and (2) a signal integration stage that incorporates these signals into a sequential recommender system through fine-tuning to align LLM behavior with recommendation objectives.

3.1. Semantic Signal Generation via LLMs

In the first stage, we utilize the reasoning capabilities of LLMs to infer semantic signals that are not explicitly present in the dataset but can be derived from item metadata such as titles, brands, and categories. This new information captures latent contextual information about user intent and item semantics. To generate contextually informative signals that enrich user-item sequences, we leverage the reasoning capabilities of LLMs through a prompting-based generation framework. Specifically, we adopt the "Let's think step by step" prompting strategy (Kojima et al., 2022), which has proven effective in eliciting structured reasoning from language models. Our approach is aligned with the Automatic Chain-of-Thought (Auto-CoT) paradigm (Zhang et al., 2022), wherein we first construct four examples using Zero-Shot-CoT, and then use them to guide the model in a few-shot setting to propose a new auxiliary signal.

Let $\mathcal{U}$ be the set of users, $\mathcal{I}$ the set of items. For a user $u \in \mathcal{U}$, their interaction sequence is $\mathbf{s}_u = [i_1, i_2, \ldots, i_t]$, where $i_k \in \mathcal{I}$. Each item i is associated with a set of structured attributes $\mathbf{x}_i = [\text{title}_i, \text{brand}_i, \text{category}_i, \cdots]$. To infer a semantic signal z_i for item i, we prompt a language model with its structured attributes:

$$z_i = \text{LLM}_\theta(\text{Prompt}(\mathbf{x}_i)), \tag{1}$$

where LLM_θ is a large language model with parameters θ, and $\text{Prompt}(\cdot)$ denotes the few-shot prompting function that formats $\mathbf{x}_i$ into textual input. The output $z_i \in \mathcal{Z}$ is a natural language phrase or sentence that captures a latent aspect of the item, such as use case or thematic intent.

The LLM (e.g., GPT-3.5) outputs a new semantic signal tailored to the item's profile and its probable user interaction context. These outputs are filtered for quality and stored as an additional item attribute to be later used during model training. The goal of this stage is to automatically derive features that encapsulate **implicit semantics**. While generating additional semantic signals through our framework, the human annotators performed a quality check to ensure that any signal inconsistent with existing metadata was discarded and not used as a new semantic feature. For example, for the "Pet Supplies" domain, the model infers a "Pet's Specific Need" signal that highlights the intended style of use. For the "Video Games" domain, the model suggests "Gameplay Experience Focus" as an augmented signal. These inferred signals extend beyond raw metadata, adding interpretability and domain-relevant nuance.

Metric	GPT	Gemini
Usefulness (Avg)	13.67	13.29
Usefulness (%)	**91.13%**	**88.60%**
Relevancy (Avg)	14.63	14.08
Relevancy (%)	**97.53%**	**93.87%**

Table 2: Human evaluation of semantic signals generated by GPT and Gemini. Average usefulness and relevancy scores over 15 samples per dataset–model pair (0, 0.5, 1).

3.2. Signal Integration and Model Alignment

In the second stage, we use the generated semantic signals to enhance sequential recommendation through two mechanisms to improve representation capacity and downstream recommendation quality: (a) input-level integration into the user-item interaction history, and (b) alignment of the sequential recommendation models or LLM's generation behavior with the recommendation objective.

Signal Integration. LLM-generated signals are incorporated into the input representation of sequential recommendation models. We first augment the item representation to include the semantic signal:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \cup \{z_i\}. \tag{2}$$

As illustrated in Table 1, we see an example for one specific product's "Pet's Specific Need" as *A balanced and nutritious staple food to support the health and growth of medium-sized fish in aquatic environments.*

We then format each item as a text-based key-value pair and construct chronological user interaction sequences that include these enriched item representations. Thus, each user's enriched interaction sequence becomes:

$$\tilde{\mathbf{s}}_u = [\tilde{\mathbf{x}}_{i_1}, \tilde{\mathbf{x}}_{i_2}, \ldots, \tilde{\mathbf{x}}_{i_t}] \tag{3}$$

This approach allows the model to consider **both explicit metadata and LLM-inferred semantics** when learning patterns from user history.

Model Alignment. Furthermore, the sequential recommenders or LLM can be fine-tuned using these augmented interaction sequences to better align its generative behavior with the needs of recommendation. To align the LM's generation behavior with recommendation utility, we fine-tune LM_θ on sequences enriched with z_i, training it to predict the next item description. The fine-tuning objective encourages the model to improve its understanding of useful semantics:

$$\mathcal{L}_{\text{gen}}(\theta) = -\sum_{u \in \mathcal{U}} \log P_\theta(i_{t+1} \mid \tilde{\mathbf{s}}_u). \tag{4}$$

By training the model on sequences that include both structured metadata and generated semantic

Dataset	Metric	ID-Text	Text-based Methods				
		S^3Rec*	ZESRec*	UniSRec*	Recformer	LaMAR (GPT-4o)	LaMAR (Gemini-1.5)
Industrial and Scientific	N@10	0.0451	0.0843	0.0862	0.1052	0.1114 (+5.89%)	0.1056 (+0.38%)
	R@10	0.0804	0.1260	0.1255	0.1479	0.1524 (+3.04%)	0.1504 (+1.69%)
	N@50	–	–	–	0.1229	0.1286 (+4.64%)	0.1238 (+0.73%)
	R@50	–	–	–	0.2288	0.2313 (+1.09%)	0.2338 (+2.19%)
	MRR	0.0392	0.0745	0.0786	0.0977	0.1044 (+6.86%)	0.0976 (-0.10%)
	AUC	–	–	–	0.7657	0.7658 (+0.01%)	0.7723 (+0.86%)
Musical Instruments	N@10	0.0797	0.0694	0.0785	0.0814	0.0838 (+2.95%)	0.0826 (+1.47%)
	R@10	0.1110	0.1078	0.1119	0.1031	0.1076 (+4.36%)	0.1064 (+3.20%)
	N@50	–	–	–	0.0952	0.0983 (+3.26%)	0.0968 (+1.68%)
	R@50	–	–	–	0.1664	0.1747 (+4.99%)	0.1722 (+3.49%)
	MRR	0.0755	0.0633	0.0740	0.0794	0.0813 (+2.39%)	0.0800 (+0.75%)
	AUC	–	–	–	0.7912	0.8058 (+1.84%)	0.7972 (+0.76%)
Arts, Crafts and Sewing	N@10	0.1026	0.0970	0.0894	0.1269	0.1282 (+1.02%)	0.1283 (+1.10%)
	R@10	0.1399	0.1349	0.1333	0.1579	0.1638 (+3.74%)	0.1646 (+4.24%)
	N@50	–	–	–	0.1415	0.1443 (+1.98%)	0.1447 (+2.26%)
	R@50	–	–	–	0.2255	0.2377 (+5.41%)	0.2399 (+6.39%)
	MRR	0.1057	0.0870	0.0798	0.1218	0.1220 (+0.16%)	0.1220 (+0.16%)
	AUC	–	–	–	0.8275	0.8361 (+1.04%)	0.8413 (+1.67%)
Office Products	N@10	0.0911	0.0865	0.0919	0.1138	0.1163 (+2.20%)	0.1173 (+3.07%)
	R@10	0.1186	0.1199	0.1262	0.1408	0.1434 (+1.85%)	0.1456 (+3.41%)
	N@50	–	–	–	0.1235	0.1262 (+2.17%)	0.1274 (+3.16%)
	R@50	–	–	–	0.1854	0.1888 (+1.83%)	0.1914 (+3.24%)
	MRR	0.0957	0.0797	0.0848	0.1084	0.1109 (+2.31%)	0.1116 (+2.95%)
	AUC	–	–	–	0.7586	0.7593 (+0.09%)	0.7618 (+0.42%)
Video Games	N@10	0.0532	0.0530	0.0580	0.0680	0.0715 (+5.15%)	0.0710 (+4.41%)
	R@10	0.0879	0.0844	0.0923	0.1039	0.1102 (+6.06%)	0.1092 (+5.10%)
	N@50	–	–	–	0.0913	0.0950 (+4.05%)	0.0948 (+3.83%)
	R@50	–	–	–	0.2120	0.2185 (+3.07%)	0.2190 (+3.30%)
	MRR	0.0500	0.0505	0.0552	0.0643	0.0669 (+4.04%)	0.0667 (+3.73%)
	AUC	–	–	–	0.8912	0.8881 (-0.35%)	0.8903 (-0.10%)
Pet Supplies	N@10	0.0742	0.0754	0.0702	0.0968	0.0992 (+2.48%)	0.0978 (+1.03%)
	R@10	0.1039	0.1018	0.0933	0.1155	0.1213 (+5.02%)	0.1212 (+4.94%)
	N@50	–	–	–	0.1049	0.1081 (+3.05%)	0.1073 (+2.29%)
	R@50	–	–	–	0.1528	0.1620 (+6.02%)	0.1648 (+7.85%)
	MRR	0.0710	0.0706	0.0650	0.0936	0.0952 (+1.71%)	0.0934 (-0.21%)
	AUC	–	–	–	0.7888	0.7959 (+0.90%)	0.7996 (+1.37%)

Table 3: Comparative analysis of LaMAR performance on semantic signals generated by `GPT-4o-mini` and `Gemini-1.5-flash` across subsets of the Amazon Review Data (Ni et al., 2019): Industrial and Scientific, Musical Instruments, Arts, Crafts and Sewing, Office Products, Video Games, and Pet Supplies. The text-based baselines refer to item representations using three features: Title, Brand, and Category. *: results are reported from Li et al. (2023). N: NDCG; R: Recall.

signals, the LLM learns to produce more relevant, context-aware outputs for future signal generation.

4. Experiments

In this section, we describe the experimental setup and aim to answer the following research question: Can integrating LLM-generated semantic signals (1) improve the performance of existing sequential recommenders and (2) enhance the recommendation capabilities of large language models?

4.1. Setup

Models. To assess the impact of the LLM-generated semantic signals, we employ RecFormer (Li et al., 2023), a state-of-the-art framework for learning language representations in sequential recommendation. RecFormer models user behavior by processing a sequence of historical items as textual input and predicting the next item based on contextual understanding. It represents items as key-value attribute pairs and encodes their sequences using a bi-directional Transformer architecture. Inspired by Longformer, this model leverages specialized embeddings for item text to effectively capture sequential patterns. We also evaluate the performance of `Llama-3.2-3B-Instruct` model (Touvron et al., 2023) and `Qwen2.5-1.5B-Instruct` model (Team, 2024) with LoRA fine-tuning.

Metrics. We use standard evaluation metrics, including NDCG@10 (Järvelin and Kekäläinen, 2002), Recall@10 (Schütze et al., 2008), NDCG@50, Recall@50, MRR (Voorhees et al., 1999), and AUC (Hanley and McNeil, 1982), and compare these values with those reported in (Li et al., 2023), which uses the RecFormer model without LLM-generated semantic signals. These metrics collectively measure both the ranking quality and retrieval accuracy of recommended items within the top positions.

Dataset	Metric	Llama		Qwen	
		Original	LaMAR (%)	Original	LaMAR (%)
Industrial and Scientific	R@10	0.7230	0.8810 (+21.85%)	0.7680	0.8560 (+11.45%)
	N@10	0.6836	0.8773 (+28.48%)	0.6996	0.8052 (+15.09%)
	MRR	0.6710	0.8761 (+30.57%)	0.6777	0.7891 (+16.44%)
Arts, Crafts and Sewing	R@10	0.7350	0.8550 (+16.33%)	0.8250	0.8530 (+3.39%)
	N@10	0.6947	0.8341 (+20.07%)	0.7696	0.8047 (+4.56%)
	MRR	0.6818	0.8273 (+21.34%)	0.7514	0.7893 (+5.05%)
Video Games	R@10	0.8080	0.9560 (+18.32%)	0.9690	0.9750 (+0.62%)
	N@10	0.8056	0.9545 (+18.48%)	0.9525	0.9666 (+1.48%)
	MRR	0.8048	0.9540 (+18.54%)	0.9472	0.9639 (+1.76%)
Pet Supplies	R@10	0.9580	0.9660 (+0.83%)	0.9180	0.9750 (+6.21%)
	N@10	0.9501	0.9608 (+1.13%)	0.8878	0.9647 (+8.66%)
	MRR	0.9475	0.9591 (+1.22%)	0.8780	0.9613 (+9.49%)
Musical Instruments	R@10	0.9370	0.9600 (+2.45%)	0.9100	0.9410 (+3.41%)
	N@10	0.9294	0.9572 (+2.99%)	0.8631	0.9149 (+6.00%)
	MRR	0.9269	0.9563 (+3.17%)	0.8481	0.9064 (+6.88%)
Office Products	R@10	0.7210	0.8990 (+24.69%)	0.6890	0.9160 (+32.98%)
	N@10	0.6912	0.8924 (+29.11%)	0.6319	0.8943 (+41.54%)
	MRR	0.6816	0.8902 (+30.60%)	0.6135	0.8873 (+44.63%)

Table 4: Comparative analysis of `LoRA` fine-tuned Llama and Qwen performance on the original (Title, Brand, and Category) versus LaMAR, which includes additional semantic signals generated by `GPT-4o-mini`. N: NDCG; R: Recall.

Implementation Details. To fine-tune *Recformer* with generated features, we set batch size to 12, learning rate to 2×10^{-5}, and weight decay to 0.01. We set *max_item_embeddings* $= 81$, *max_token_num* $= 256$, and *max_attr_num* $= 4$. For fine-tuning with the original feature, we keep the default configuration of `Recformer` and only modify *max_token_num* to 256. Fine-tuning was conducted on $4\times$ NVIDIA A5000 GPU (24GB memory). We fine-tune the `Llama-3.2-3B-Instruct` model using a lightweight adapter-based approach, LoRA (Low-Rank Adaptation) (Hu et al., 2022), which introduces trainable low-rank decomposition matrices into the model's layers while keeping the original weight matrices frozen. The LoRA configuration included a rank of $r = 16$, $\alpha = 32$, and a dropout rate of 0.1. The target modules for adaptation are q_proj, k_proj, and v_proj. The learning rate is set to 2×10^{-4} and fine-tuning was conducted on $2\times$ NVIDIA A5000 GPU (24GB memory). We configure *per_device_train_batch_size* to 2 and *gradient_accumulation_steps* to 20, resulting in an effective batch size of 80. The same experimental configuration was applied to the `Qwen2.5-1.5B-Instruct` model, while using one GPU and a learning rate of 10^{-4}.

4.2. Semantic Signal Generation

In this work, we apply the signal generation framework across six categories from the Amazon review dataset (Ni et al., 2019): Industrial and Scientific, Musical Instruments, Arts, Crafts and Sewing, Office Products, Video Games, and Pet Supplies. For each dataset, we apply the signal generation pipeline using `GPT-4o-mini` (OpenAI, 2024) and `Gemini-1.5-flash-002` (Team et al., 2024) with a consistent 3-shot prompting configuration. Table 1 outlines the semantic signal generated for each domain and detailed description of the generated signals. Specifically, it summarizes the nature of the semantic signal introduced per category, such as "Primary Use Case" for Industrial products or "Gameplay Experience Focus" for Video Games, all automatically generated by the LLM during the prompting stage. It also provides concrete instances where these signals are added to the item profile, showing how they complement the existing metadata. These enriched sequences are used to train and evaluate the RecFormer model.

These newly generated signals are integrated as a fourth attribute in addition to the standard triplet of Title, Brand, and Category. In our ablation studies, we also ablate the prompt and number of generated semantic signals in Section 5.1. To evaluate the quality of LLM-generated semantic signals, we conduct a small-scale human study. The annotators, all graduate students with prior experience in recommender systems, rated 180 samples (15 per dataset per model) on two criteria (scored as 0, 0.5, or 1): (1) Relevancy: factual or contextual alignment with the input, and (2) Usefulness: added informative value beyond basic fields. The results in Table 2 indicate that the majority of generated signals were both relevant and informative, suggesting that the LLM-generated semantic signals generally provide useful and contextually appropriate information for the task.

Metric	Original	LaMAR	Improv.	LaMAR (prompt variant)	Improv.	LaMAR (signal variant)	Improv.
N@10	0.1052	0.1114	**+5.89%**	0.1080	+2.66%	0.1086	+3.23%
R@10	0.1479	0.1524	**+3.04%**	0.1520	+2.77%	0.1477	-0.14%
N@50	0.1229	0.1286	**+4.64%**	0.1255	+2.12%	0.1263	+2.77%
R@50	0.2288	0.2313	+1.09%	0.2324	**+1.57%**	0.2290	+0.09%
MRR	0.0977	0.1044	**+6.86%**	0.1000	+2.35%	0.1022	+4.61%
AUC	0.7657	0.7658	**+0.01%**	0.7637	-0.26%	0.7624	-0.43%

Table 5: **Comparison of LaMAR variants on the Scientific dataset.** We report performance using the original features (Title, Brand, Category), the standard LaMAR signal, and two variants: prompt-based (different prompting method) and signal-based (multiple signals). All semantic signals are generated using `GPT-4o-mini`.

4.3. Can Semantic Signal Integration Enhance Sequential Recommenders?

As the results show in Table 3, we observe an improvement in performance across all domains using both language models, highlighting the effectiveness of our proposed framework in enhancing sequential recommendation performance.

For instance, for the Video Games dataset, the original model achieved an NDCG@10 of 0.0680 and a Recall@10 of 0.1039. After incorporating new signals through our framework, the model's performance improved: with GPT-generated signals, NDCG@10 increased to 0.0715 and Recall@10 to 0.1102, representing improvements of 5.15% and 6.06%, respectively; with Gemini-generated signals, NDCG@10 increased to 0.0710 and Recall@10 to 0.1092, corresponding to improvements of 4.41% and 5.10%, respectively. These results show that LaMAR enables the model to rank more relevant items higher and find more relevant items overall, improving performance regardless of whether GPT or Gemini is used.

4.4. Can Semantic Signal Integration Improve LLM-Based Recommendation?

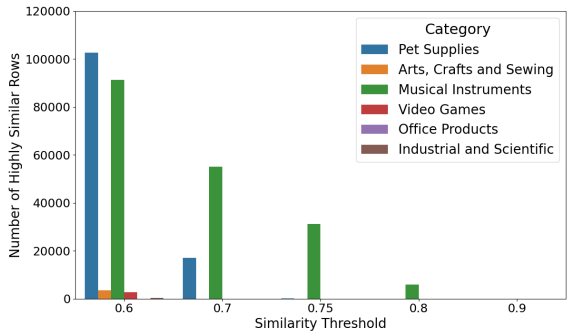
To evaluate the effectiveness of data augmentation, we utilize two datasets: one with augmented data generated by GPT-4o-mini and another containing only the original data (where an item is represented with three features: Title, Brand, and Category), and compare their performance on a sequential recommendation task. For this purpose, we construct prompts consisting of five chronological user interactions, followed by a candidate pool of 20 randomly selected items along with the ground truth item. The order of all candidate items in each prompt was randomized to avoid positional bias during recommendation. We then prompt the model to recommend the next item from this pool. For our experiments, we adopt the `Llama-3.2-3B-Instruct` model (Touvron et al., 2023) and `Qwen2.5-1.5B-Instruct` model (Team, 2024) as the base LLMs and used $5{,}000$ prompts for fine-tuning. The results are presented in Table 4. The results confirm the effectiveness of our proposed framework; however, we observe that for datasets where the generated features are highly similar, the model shows only slight or no improvement. This indicates when using LLMs for recommendation systems, similar items can limit the benefit of data augmentation. Adding highly similar information can even lower the performance of the model, as it is shown in Table 4 for "Pet Supplies" and "Musical Instruments" datasets, where the performance gains were minimal when similar features were added. A detailed analysis of signal diversity is provided in Section 5.2. To assess stability, we repeated each experiment three times and report the mean and standard deviation in Appendix A.6. The results show consistent gains across all domains, suggesting that the improvements result from the added semantic signals rather than random initialization.

5. Additional Results and Analysis

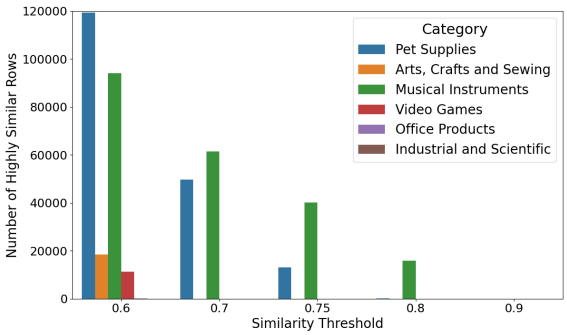
5.1. Ablation on Prompts and Signal Numbers

We further evaluate LaMAR's robustness on the Scientific dataset by examining two variants: an alternative prompting strategy and an expanded multi-signal setup. Specifically, we investigate whether modifying the prompt design or generating additional semantic signals improves recommendation performance. In the prompt variant, we modify the input format by including the dataset name, available features, and several random examples, prompting GPT-4o-mini to generate a new semantic signal. In the signal variant, we apply the standard LaMAR prompting strategy iteratively, generating and integrating a second signal.

As shown in Table 5, both variants improve over the original feature set (Title, Brand, Category). However, the standard LaMAR configuration

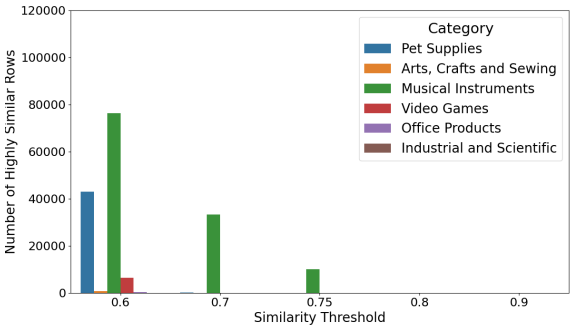


(a) Semantic similarity analysis for `GPT-4o-mini`.

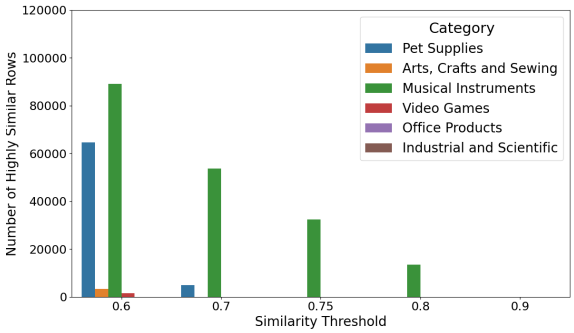


(b) Semantic similarity analysis for `Gemini-1.5-Flash`.

Figure 2: Semantic similarity analysis across thresholds for signals generated by `GPT-4o-mini` and `Gemini-1.5-Flash`, using `multi-qa-mpnet-base-cos-v1`.



(a) Semantic similarity analysis for `GPT-4o-mini`.



(b) Semantic similarity analysis for `Gemini-1.5-Flash`.

Figure 3: Semantic similarity analysis across thresholds for signals generated by `GPT-4o-mini` and `Gemini-1.5-Flash`, using `all-MiniLM-L6-v2`.

yields the highest overall performance, with the largest gains in NDCG@10 (+5.89%), Recall@10 (+3.04%), and MRR (+6.86%). The prompt variant shows slightly lower effectiveness, with performance drops of 2.66%, 2.77%, and 2.35% in those same metrics, respectively, compared to LaMAR. The signal variant, while introducing a second semantic signal, does not lead to further gains and even slightly underperforms in Recall@10 and AUC, suggesting potential redundancy. These findings highlight that careful prompt design is critical, and that adding more semantic signals does not always translate to better recommendation performance.

5.2. Signal Diversity Analysis

To explore the novelty of the generated features for each dataset, we compare the semantic similarity, focusing on meaning and context rather than surface-level word overlap. We embed each sentence into a high-dimensional vector space, and compute the cosine similarity between pairs of vectors. We use two embedding models: a general-purpose model, `all-MiniLM-L6-v2`, and a semantically focused model, `multi-qa-mpnet-base-cos-v1`. To identify similar texts, we apply five cosine similarity thresholds: 0.6, 0.7, 0.75, 0.8, and 0.9. We categorize a text as "highly similar" in a given threshold when it is similar to more than $0.1\times$ the length of the dataset.

Figures 2 and 3 represent the number of highly similar rows in each dataset for a feature generated by `GPT-4o-mini` and `Gemini-1.5-Flash`. The plot reveals that for features generated by both models, the semantic model detects a high number of similar rows at lower thresholds ($0.6 - 0.7$) especially for categories like Pet Supplies and Musical Instruments, indicating that the LLM may have reused similar phrasing or templates in these domains – for example, recurring patterns involving common pet types or limited music genres. This suggests that during generation, certain categories are more prone to semantic repetition. In addition, features generated by `Gemini-1.5-Flash` show higher similarity compared to those generated by `GPT-4o-mini` when comparing general and semantic similarity. Importantly, when we look beyond the 0.7 threshold and exclude the Pet Supplies and Mu-

Metric	Recformer	LaMAR (GPT-4o)	Improv. (%)
NDCG@10	0.0384	0.0453	+17.97%
Recall@10	0.0774	0.0926	+19.64%
NDCG@50	0.0710	0.0782	+10.14%
Recall@50	0.2288	0.2444	+6.82%
MRR	0.0367	0.0409	+11.44%
AUC	0.8264	0.8241	-0.27%

Table 6: Comparing Recformer and LaMAR.

sical Instruments categories, the number of highly similar rows drops dramatically. This means that in most other domains, each generated feature is semantically distinct from over 90% of the dataset, demonstrating strong content diversity. This provides confidence that LLMs can generate varied and original outputs in less templated categories.

5.3. Generalization to a Different Domain

To assess the generalizability of our approach beyond Amazon product categories, we conducted additional experiments on the MovieLens-1M dataset. In this setting, we used the movie title and genre as base features. Since the dataset does not include storyline information, we enriched it by incorporating storyline data obtained through web scraping. Furthermore, we introduced a new semantic signal, "how the movie ends", generated by prompting `GPT-4o-mini` via LaMAR to enrich item representations. Table 6 summarizes the results, showing consistent improvements across multiple metrics. These results demonstrate that LaMAR generalizes effectively across domains, and the proposed semantic signal contributes meaningful improvements even outside the e-commerce setting.

6. Conclusion

We propose LaMAR, a framework that uses LLMs to generate semantic signals for augmenting sequential recommendation via few-shot prompting, eliminating the need for manual feature engineering. Experiments across domains show consistent performance gains, demonstrating the effectiveness of LLMs as scalable, data-centric semantic augmenters.

7. Limitation

We note that in real-world scenarios, the ground-truth item may not always be present in the candidate pool. Consequently, reported metrics such as Recall and NDCG should be interpreted as an upper bound, though the setup still allows a controlled comparison between augmented and original item representations. Additionally, due to computational constraints, we limited the candidate pool to 21 items. However, in real-world scenarios, the number of available items is typically much larger.

Also, we note that the results were obtained using a proprietary LLM (e.g., GPT-4o-mini), whose architecture and training data may evolve over time, and the exact reproducibility of these results may be affected. However, our experiments demonstrate the effectiveness of the approach in leveraging LLM-generated semantic signals, and the methodology is generalizable to other LLMs or future versions.

8. Acknowledgement

The authors have no acknowledgements to declare.

9. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Millennium Bismay, Xiangjue Dong, and James Caverlee. 2025. ReasoningRec: Bridging personalized recommendations and human-interpretable explanations through LLM reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8132–8148, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander Brinkmann, Roee Shraga, Reng Chiz Der, and Christian Bizer. 2023. Product information extraction using chatgpt. *arXiv preprint arXiv:2306.14921*.

Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 108–116.

Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318*.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Ruining He, Wang-Cheng Kang, and Julian McAuley. 2017. Translation-based recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 161–169.

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 585–593.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Feiran Huang, Yuanchen Bei, Zhenghang Yang, Junyi Jiang, Hao Chen, Qijie Shen, Senzhang Wang, Fakhri Karray, and Philip S Yu. 2025. Large language model simulator for cold-start recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 261–270.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1258–1267.

Yueqing Liang, Liangwei Yang, Chen Wang, Xiongxiao Xu, Philip S Yu, and Kai Shu. 2024. Taxonomy-guided zero-shot recommendations with llms. *arXiv preprint arXiv:2406.14043*.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, et al. 2023. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*.

Jianghao Lin, Rong Shan, Chenxu Zhu, Kounianhua Du, Bo Chen, Shigang Quan, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation. In *Proceedings of the ACM on Web Conference 2024*, pages 3497–3508.

Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. 2024a. Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*.

Fan Liu, Yaqi Liu, Huilin Chen, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. 2025. Understanding before recommendation: Semantic aspect-aware review exploitation via large language models. *ACM Transactions on Information Systems*, 43(2):1–26.

Qidong Liu, Xiangyu Zhao, Yuhao Wang, Yejing Wang, Zijian Zhang, Yuqi Sun, Xiang Li, Maolin Wang, Pengyue Jia, Chong Chen, et al. 2024b. Large language model enhanced recommender systems: A survey. *arXiv preprint arXiv:2412.13432*.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024c. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 452–461.

Sichun Luo, Jiansheng Wang, Aojun Zhou, Li Ma, and Linqi Song. 2024. Large language models augmented rating prediction in recommender system. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7960–7964. IEEE.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Yankun Ren, Zhongde Chen, Xinxing Yang, Longfei Li, Cong Jiang, Lei Cheng, Bo Zhang, Linjian

Mo, and Jun Zhou. 2024. Enhancing sequential recommenders with augmented knowledge from aligned large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 345–354.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, Ed Chi, and Xinyang Yi. 2024. Better generalization with semantic ids: A case study in ranking for recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 1039–1044, New York, NY, USA. Association for Computing Machinery.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Jianling Wang, Haokai Lu, James Caverlee, Ed H Chi, and Minmin Chen. 2024a. Large language models as data augmenters for cold-start item recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 726–729.

Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Zhang, Qing Cui, et al. 2024b. Llmrg: Improving recommendations through large language model reasoning graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19189–19196.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 12–22.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 1259–1273. IEEE.

Yuki Yada and Hayato Yamana. 2024. News recommendation with category description by a large language model. *arXiv preprint arXiv:2405.13007*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1893–1902.

Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32.

10. Language Resource References

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

OpenAI. 2024. Gpt-4o-mini. `https://platform.openai.com/docs/models/gpt-4o-mini`.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

A. More Implementation Details

A.1. Prompts

An example of prompt employed for proposing new semantic signal and for generating the detailed description of LLM-generated semantic signals are shown in Figures 4 and 5 respectfully. The prompts used to generate the fifth feature described in 5.1 are also shown in 6, 7, and 8.

A.2. Metrics

We evaluate the effectiveness of our framework using standard metrics commonly adopted in recommendation research, including NDCG@10 (Järvelin and Kekäläinen, 2002), Recall@10 (Schütze et al., 2008), NDCG@50, Recall@50, MRR (Voorhees et al., 1999), and AUC (Hanley and McNeil, 1982; He et al., 2017). These metrics collectively measure both the ranking quality and retrieval accuracy of recommended items within the top positions. NDCG (Normalized Discounted Cumulative Gain) assesses the position-weighted relevance of recommended items, while Recall@K evaluates the proportion of relevant items retrieved among the top-K. MRR (Mean Reciprocal Rank) emphasizes the rank of the first correct item, and AUC (Area Under the ROC Curve) captures the overall discriminative ability of the model across all thresholds. We report these values for all experimental configurations and directly compare them to the baseline results reported in (Li et al., 2023), which uses the RecFormer model without LLM-generated features. This comparison allows us to quantify the contribution of our semantic augmentation framework across diverse datasets and evaluation dimensions.

A.3. Models

Table 7 shows the models and their licenses.

Model	Link	License
Llama-3.2-3B-Instruct	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct	llama3.2
GPT-4o-mini	https://platform.openai.com/docs/models/gpt-4o-mini	OpenAI
Gemini-1.5-flash	https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-flash	Google
Qwen2.5-1.5B-Instruct	https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct	Apache License 2.0

Table 7: Models.

A.4. Data Statistics

Table 8 shows the statistics of the dataset.

Category	Users	Items	Interactions
Industrial and Scientific	11,041	5,327	76,896
Musical Instruments	27,530	10,611	231,312
Arts, Crafts and Sewing	56,210	22,855	492,492
Office Products	101,501	27,932	798,914
Video Games	11,036	15,402	100,255
Pet Supplies	47,569	37,970	420,662

Table 8: Data statistics.

A.5. Data Preprocessing

For RecFormer, we follow the data processing procedure described in (Li et al., 2023). Specifically, we use the datasets provided by the original data sources and filter out items with missing titles. User interactions are grouped by user and sorted in ascending order by timestamp to form sequential input. For each item, we extract key attributes—such as title, categories, and brand—and represent them as key-value pairs. For fine-tuning LLaMA, each training example is constructed from five chronologically ordered items from a user's interaction history, enriched with their corresponding semantic features, with the next item in the sequence used as the ground-truth target.

A.6. Stability Analysis

To assess robustness, we repeated experiments three times with different random seeds using the `Qwen2.5-1.5B-Instruct` model (Team, 2024), fine-tuned with `LoRA`, and report the mean $\pm$ standard deviation in Table 9. The results show consistently low variance across runs, indicating stable optimization behavior. Moreover, LaMAR consistently outperforms the 3-feature baseline across all domains, suggesting that the improvements arise from the added semantic signals rather than random initialization.

A.7. Baseline Analysis

We employ three types of methods as our baseline similar to (Li et al., 2023), including approaches that use item IDs as primary inputs while incorporating item text as auxiliary information ($S^3 - Rec$); and approaches have exclusively item texts as input (ZESRec, UniSRec, Recformer). $S^3 - Rec$ (Zhou et al., 2020) improves sequential recommendation by leveraging data correlations for self-supervision signals generation and data representations improvement. ZESRec (Ding et al., 2021) is trained on an old dataset and generalizes to a new one with no overlapping users or items. UniSRec (Hou et al., 2022) utilizes a lightweight MoE-based module to integrate textual item representations.

We also compare our framework against TAXREC (Liang et al., 2024), a state-of-the-art taxonomy-guided augmentation approach. The re-

Model	Metric	Industrial and Scientific	Musical Instruments	Arts, Crafts and Sewing	Office Products	Video Games	Pet Supplies
Original	Recall@10	$0.7900_{\pm 0.0161}$	$0.9017_{\pm 0.0065}$	$0.8260_{\pm 0.0014}$	$0.6250_{\pm 0.0453}$	$0.8753_{\pm 0.0663}$	$0.9370_{\pm 0.0134}$
	NDCG@10	$0.7359_{\pm 0.0258}$	$0.8560_{\pm 0.0052}$	$0.7679_{\pm 0.0021}$	$0.5540_{\pm 0.0551}$	$0.8292_{\pm 0.0872}$	$0.9087_{\pm 0.0150}$
	MRR	$0.7186_{\pm 0.0291}$	$0.8413_{\pm 0.0049}$	$0.7489_{\pm 0.0031}$	$0.5313_{\pm 0.0581}$	$0.8143_{\pm 0.0940}$	$0.8994_{\pm 0.0155}$
LaMAR	Recall@10	$\mathbf{0.8597_{\pm 0.0026}}$	$\mathbf{0.9423_{\pm 0.0019}}$	$\mathbf{0.8547_{\pm 0.0017}}$	$\mathbf{0.9173_{\pm 0.0009}}$	$\mathbf{0.9737_{\pm 0.0012}}$	$\mathbf{0.9763_{\pm 0.0012}}$
	NDCG@10	$\mathbf{0.8036_{\pm 0.0012}}$	$\mathbf{0.9161_{\pm 0.0027}}$	$\mathbf{0.8035_{\pm 0.0014}}$	$\mathbf{0.8955_{\pm 0.0009}}$	$\mathbf{0.9661_{\pm 0.0007}}$	$\mathbf{0.9635_{\pm 0.0008}}$
	MRR	$\mathbf{0.7857_{\pm 0.0024}}$	$\mathbf{0.9076_{\pm 0.0031}}$	$\mathbf{0.7871_{\pm 0.0019}}$	$\mathbf{0.8885_{\pm 0.0010}}$	$\mathbf{0.9636_{\pm 0.0010}}$	$\mathbf{0.9593_{\pm 0.0014}}$

Table 9: Robustness comparison between the original (Title, Brand, and Category) and LaMAR, which includes additional semantic signals generated by `GPT-4o-mini`. Results are obtained using the `Qwen2.5-1.5B-Instruct` model fine-tuned with `LoRA`. N: NDCG; R: Recall.

Method	Metric	Industrial and Scientific	Musical Instruments	Arts, Crafts and Sewing	Office Products	Video Games	Pet Supplies
TAXREC (GPT-4o mini)	N@10	0.0021	0.0009	0.0007	0.0001	0.0034	0.0003
	R@10	0.0046	0.0016	0.0017	0.0004	0.0061	0.0004
	N@50	0.0050	0.0026	0.0019	0.0005	0.0081	0.0006
	R@50	0.0182	0.0099	0.0074	0.0020	0.0285	0.0017
	MRR	0.0029	0.0017	0.0011	0.0004	0.0048	0.0005
	AUC	0.5341	0.5831	0.5919	0.5085	0.6219	0.5519
LaMAR (GPT-4o mini)	N@10	0.1114	0.0838	0.1282	0.1163	0.0715	0.0992
	R@10	0.1524	0.1076	0.1638	0.1434	0.1102	0.1213
	N@50	0.1286	0.0983	0.1443	0.1262	0.0950	0.1081
	R@50	0.2313	0.1747	0.2377	0.1888	0.2185	0.1620
	MRR	0.1044	0.0813	0.1220	0.1109	0.0669	0.0952
	AUC	0.7658	0.8058	0.8361	0.7593	0.8881	0.7959

Table 10: Performance comparison between TAXREC (Liang et al., 2024) and LaMAR (GPT-4o mini) across six Amazon domains. The LaMAR results are reproduced from Table 3. N: NDCG; R: Recall.

sults are reported in Table 10. To ensure a fair comparison, we implement TAXREC using a five-feature taxonomy, aligning it closely with the original setting and keeping the feature dimensionality comparable to our framework (which operates with four semantic signals). This ensures that performance differences are not driven by feature count discrepancies. In addition, we consider the entire user interaction history when constructing user representations, whereas the original TAXREC implementation focused on the 10 most recent interactions. TAXREC yields lower performance on the Amazon product domains compared to both its originally reported results on MovieLens and Book-Crossing, as well as to LaMAR. We attribute this performance gap to several structural differences in the evaluation settings.

First, the scale of the candidate pool differs significantly across these benchmarks. TAXREC was originally evaluated on MovieLens-100K (Harper and Konstan, 2015), which contains 1,682 movies, and Book-Crossing (Ziegler et al., 2005), which includes 4,389 items. In contrast, the Amazon domains considered in our experiments are substantially larger, ranging from 5,000 to 38,000 unique items (Table 8). Since TAXREC ranks items by counting feature-value overlaps within a fixed five-feature taxonomy, the maximum possible score is five. In larger catalogs, this naturally leads to many items sharing identical scores. As the candidate set grows, distinguishing the target item among these tied candidates becomes increasingly challenging, which in turn affects Recall@k and NDCG@k.

Second, the nature of the taxonomy varies by domain. Movie taxonomies are inherently more discriminative than product taxonomies. The original TAXREC movie taxonomy includes features such as genre, IMDb rating, and Rotten Tomatoes score; attributes that are highly structured, semantically meaningful, and well-represented in LLM pretraining data. In contrast, product domains lack similarly standardized and universally recognized attributes, reducing the discriminative power of taxonomy-based matching. Moreover, product titles are typically noisier and less semantically informative than movie titles, which may limit the effectiveness of categorical feature overlap.

These results suggest that while taxonomy-based approaches are effective in structured, moderate-scale settings, high-cardinality and diverse product catalogs benefit from more scalable semantic representations. By leveraging richer LLM-generated signals, LaMAR enables finer-grained differentiation across items even in high-cardinality domains. As shown in Table 10, LaMAR consistently achieves strong improvements across all tested domains, demonstrating the effectiveness of integrating LLM-derived semantic information for large-scale recommendation settings.

Your task is introducing a new feature that can improve sequential recommendation task in a given dataset. The number of words must be between 30 and 50 words.

Input:
task type: Next movie recommendation to user based on pervious interaction.
Dataset: MovieLens 1M
Available feature: Title, Rating, Genre, Storyline.
Output:
Let's think step by step. A user is lean toward seeing movie based on the feeling experiences at the end of movie. Therefore, how the movie ends can have impact on next movie the user selects to watch.
Recommended feature: How movie ends.

Input:
task type: Next beauty product recommendation to user based on pervious purchase history.
Dataset: Amazon All Beauty
Available feature: Title, User_Rating, Average_Rating, Description.
Output:
Let's think step by step. A user's preference for beauty products may depend on the product's key benefits or unique selling points. Therefore, highlighting the main benefit can influence the next purchase decision.
Recommended feature: Key Benefit.

Input:
task type: Next video game recommended to user based on pervious game history.
Dataset: Amazon Video Games
Available feature: Title, Brand, Rating, Summary of review.
Output:
Let's think step by step. A user's choice for the next game may depend on the gameplay experience, such as its genre or game mechanics. Therefore, a new impactful feature is the game style or genre.
Recommended feature: Game Style.

Input:
task type: Next book recommended to user based on interaction.
Dataset: Amazon Books
Available feature: Title, Description, Author, Summary of review
Output:
Let's think step by step. A user's next book choice may depend on the central theme or primary emotion evoked by the book. Therefore, a new impactful feature is the main theme or emotional tone of the book.
Recommended feature: Main Theme/Emotion.

Input:
task type: Next musical instruments recommended to user based on interaction.
Dataset: Amazon Musical Instruments
Available feature: Title, Brand, Category
Output:
Let's think step by step.

A user's choice for the next musical instrument may depend on the intended music genre or playing skill level, as these factors shape usability and appeal. Therefore, a new impactful feature is the primary music genre or skill suitability.
Recommended feature: Music Genre/Skill Suitability.

Figure 4: An example of prompt used for proposing a new relevant signal.

You are given the information of an office product including its title, brand, and category as input. The output must be the Product Function and should have between 10 to 30 words. Do not include "Product Function:" in your answer.

Input:
Title: Alligator Leather-Look Organizer Black XL Book and Bible Cover
Brand: Visit Amazon's Zondervan Page
Category: Office Products Office & School Supplies Binders & Binding Systems

Output:
Product Function: Keeps books, Bibles, and documents securely organized and protected with a stylish, durable alligator leather-look exterior, offering a professional and functional storage solution.

Input:
Title: LIFEPAC 2nd Grade Language Arts Set of 10 LIFEPACs
Brand: Alpha Omega Publications
Category: Office Products Office & School Supplies Education & Crafts Special Education Supplies

Output:
Product Function: Provides structured language arts curriculum for 2nd grade, supporting reading, writing, and comprehension skills through a series of engaging and comprehensive workbooks.

Input:
Title: Big Judy Clock Bulletin Board Set
Brand: Carson-Dellosa
Category: Office Products Office & School Supplies Education & Crafts Arts & Crafts Supplies Classroom Decorations

Output:
Product Function: Combines a functional clock with a bulletin board for organizing schedules, announcements, and classroom materials, enhancing classroom management and timekeeping.

Input:
Title: Lucie Summers Eco Writer's Notebook
Brand: Galison
Category: Office Products Office & School Supplies Paper

Output:
Product Function:

Encourages creativity and note-taking with environmentally friendly materials, offering ample space for writing, journaling, and brainstorming ideas in a stylish notebook design.

Figure 5: An example of prompt used for generating the proposed signal.

You are a trusted assistant. Your task is introducing a new feature that can improve sequential recommendation task in a given dataset. The number of words must be between 30 and 50 words.

Dataset: Amazon Industrial and Scientific
Available feature: Title, Brand, Category

Here is some items in the dataset:
Mastech MS5900 3 Motor Phase Rotation Indicator Meter Sequence Tester LED Field
5429076 PT# -35053 Bib-Eze Disposable Bib Holders 250/Pk by, DUX Dental VanR Cadco Clive -5429076
Crest 3d Tp Vivid Wht Min Size 4.1z Crest 3d White Advanced Vivid Glamorous White Tooth Paste Vibrant Mint 4.1
Hub City PB250URX1/2 Pillow Block Mounted Bearing, Normal Duty, Low Shaft Height, Relube, Setscrew Locking Collar, Narrow Inner Race, Cast Iron Housing, 1/2" Bore, 1.32" Length Through Bore, 1.063" Base To Height
Accuform MCHL224VA Aluminum Safety Sign, Legend "DANGER DIESEL FUEL", 7" Length x 10" Width, Red/Black on White
"Visitor Parking" Sign with Right Arrow By SmartSign | 12" x 18" 3M Engineer Grade Reflective Aluminum
Accuform SBMEXT566VP Plastic Spanish Bilingual Sign, Legend "EMERGENCY EXIT ONLY SECURITY ALARM WILL SOUND IF DOOR IS OPENED/SALIDA DE EMERGENCIA SOLAMENTE LA ALARMA DE SEGURIDAD SONARA AL ABRIRSE LA PUERTA", 14" Length x 10" Width x 0.055" Thickness, Red on White
Online Plastic Supply White PVC Expanded Sheet 5/32" (4mm) x 12" x 12" (8 Pack)
2 Pole Blade Relay Base Fits RH2B Relays Refrigeration Machine Accessories kits
Monroe 555031 Gas-Magnum 65 Shock Absorber

Now, introduce another feature for the given dataset.
Let’s think step by step.

Let’s think step by step. In industrial and scientific products, users often rely on specific product characteristics to make informed decisions. One key aspect is **Product Dimensions/Size**, which helps users assess the fit and suitability of the product for their intended use.

Recommended feature: Product Dimensions/Size.

Figure 6: An example of the prompt used for generating through prompt variant.

Your task is introducing a new feature that can improve sequential recommendation task in a given dataset. The number of words must be between 30 and 50 words.

Input:
task type: Next movie recommendation to user based on pervious interaction.
Dataset: MovieLens 1M
Available feature: Title, Rating, Genre, Storyline.
Output:
Let's think step by step. A user is lean toward seeing movie based on the feeling experiences at the end of movie. Therefore, how the movie ends can have impact on next movie the user selects to watch.
Recommended feature: How movie ends.

Input:
task type: Next beauty product recommendation to user based on pervious purchase history.
Dataset: Amazon All Beauty
Available feature: Title, User_Rating, Average_Rating, Description.
Output:
Let's think step by step. A user's preference for beauty products may depend on the product's key benefits or unique selling points. Therefore, highlighting the main benefit can influence the next purchase decision.
Recommended feature: Key Benefit.

Input:
task type: Next video game recommended to user based on pervious game history.
Dataset: Amazon Video Games
Available feature: Title, Brand, Rating, Summary of review.
Output:
Let's think step by step. A user's choice for the next game may depend on the gameplay experience, such as its genre or game mechanics. Therefore, a new impactful feature is the game style or genre.
Recommended feature: Game Style.

Input:
task type: Next book recommended to user based on interaction.
Dataset: Amazon Books
Available feature: Title, Description, Author, Summary of review
Output:
Let's think step by step. A user's next book choice may depend on the central theme or primary emotion evoked by the book. Therefore, a new impactful feature is the main theme or emotional tone of the book.
Recommended feature: Main Theme/Emotion.

Input:
task type: Next industrial and scientific product recommended to user based on interaction.
Dataset: Amazon Industrial and Scientific
Available feature: Title, Brand, Category
Output:
Let's think step by step.

A user's next choice of industrial and scientific product may depend on its specific functionality or technical specification, which fulfills a precise need. Therefore, a new impactful feature is the product's key functionality or technical specification.
Recommended feature: Key Functionality/Technical Specification.

Figure 7: Example prompt used for proposing the fifth semantic signal – first part.

Thank you for your suggestions. You have suggested following features for the next feature:
1- Key Functionality
2- Technical Specification
3- Application/use case
4- Compatibility

Compatibility means a state in which two things are able to exist or occur together without problems or conflict.
Application means practical use or relevance.
Functionality means the purpose that something is designed or expected to fulfill.
Technical Specification means a detailed document that outlines the requirements and features needed for a product or system to work properly.

Here is some items in the dataset:
Mastech MS5900 3 Motor Phase Rotation Indicator Meter Sequence Tester LED Field
5429076 PT# -35053 Bib-Eze Disposable Bib Holders 250/Pk by, DUX Dental VanR Cadco Clive - 5429076
Crest 3d Tp Vivid Wht Min Size 4.1z Crest 3d White Advanced Vivid Glamorous White Tooth Paste Vibrant Mint 4.1
Hub City PB250URX1/2 Pillow Block Mounted Bearing, Normal Duty, Low Shaft Height, Relube, Setscrew Locking Collar, Narrow Inner Race, Cast Iron Housing, 1/2" Bore, 1.32" Length Through Bore, 1.063" Base To Height
Accuform MCHL224VA Aluminum Safety Sign, Legend "DANGER DIESEL FUEL", 7" Length x 10" Width, Red/Black on White
"Visitor Parking" Sign with Right Arrow By SmartSign | 12" x 18" 3M Engineer Grade Reflective Aluminum
Accuform SBMEXT566VP Plastic Spanish Bilingual Sign, Legend "EMERGENCY EXIT ONLY SECURITY ALARM WILL SOUND IF DOOR IS OPENED/SALIDA DE EMERGENCIA SOLAMENTE LA ALARMA DE SEGURIDAD SONARA AL ABRIRSE LA PUERTA", 14" Length x 10" Width x 0.055" Thickness, Red on White
Online Plastic Supply White PVC Expanded Sheet 5/32" (4mm) x 12" x 12" (8 Pack)
2 Pole Blade Relay Base Fits RH2B Relays Refrigeration Machine Accessories kits
Monroe 555031 Gas-Magnum 65 Shock Absorber

We decided to have Primary Use Case as the fourth feature.
Now, introduce another feature for the given dataset.

Given the dataset, an additional useful feature could be **Brand Reputation**. The brand's reputation often influences purchasing decisions, especially for industrial and scientific products. Users may prefer products from well-known, trusted brands due to their reliability, durability, and performance standards.

Recommended feature: Brand Reputation.

Figure 8: Example prompt used for proposing the fifth semantic signal – second part.