Large Language Model Augmented Narrative Driven Recommendations

SHESHERA MYSORE, University of Massachusetts Amherst, USA

ANDREW MCCALLUM, University of Massachusetts Amherst, USA

HAMED ZAMANI, University of Massachusetts Amherst, USA

Narrative-driven recommendation (NDR) presents an information access problem where users solicit recommendations with verbose descriptions of their preferences and context, for example, travelers soliciting recommendations for points of interest while describing their likes/dislikes and travel circumstances. These requests are increasingly important with the rise of natural language-based conversational interfaces for search and recommendation systems. However, NDR lacks abundant training data for models, and current platforms commonly do not support these requests. Fortunately, classical user-item interaction datasets contain rich textual data, e.g., reviews, which often describe user preferences and context – this may be used to bootstrap training for NDR models. In this work, we explore using large language models (LLMs) for data augmentation to train NDR models. We use LLMs for authoring synthetic narrative queries from user-item interactions with few-shot prompting and train retrieval models for NDR on synthetic queries and user-item interaction data. Our experiments demonstrate that this is an effective strategy for training small-parameter retrieval models that outperform other retrieval and LLM baselines for narrative-driven recommendation.

1 INTRODUCTION

Recommender systems personalized to users are an important component of several industry-scale platforms [16, 17, 46]. These systems function by inferring users' interests from their prior interactions on the platform and making recommendations based on these inferred interests. While recommendations based on historical interactions are effective, users soliciting recommendations often start with a vague idea about their desired target items or may desire recommendations depending on the context of use, often missing in historical interaction data (Figure 1). In these scenarios, it is common for users to solicit recommendations through long-form narrative queries describing their broad interests and context. Information access tasks like these have been studied as narrative-driven recommendations (NDR) for items ranging from books [5] and movies [18], to points of interest [1]. Bogers and Koolen [5] note these narrative requests to be common on discussion forums and several subreddits[1], but, there is a lack of support for these complex natural language queries in current recommenders.

---

[1] r/MovieSuggestions, r/booksuggestions, r/Animesuggest

Looking for restaurants and food in Rotterdam, NL
Hi all - my father and I will be visiting your fair city next week, and we're looking for a place to have dinner on a Sunday night. Are there any places that you would recommend? We like tasty food, good beer, and a non-stuffy atmosphere. Considerations: Open on Sunday (I gather this is somewhat limiting?) Our hotel is right near the Leuvehaven metro stop, and walking distance would be nice Ideally not Asian(-inspired) food - we're from southeast Asia and would rather try something different. Totally up for everything else though! Thank you!.

Fig. 1. An example narrative query soliciting point of interest recommendations. The query describes the users preferences and the context of their request.

3 few-shot examples:

A user likes these recommendations: Mission BBQ in Deptford Township, Keswick Tavern in Glenside …
The user wrote these reviews: This place is always packed! Keswick Tavern is a go to if you're looking for great food and large alcohol/ beer selection …
In response to the request on Reddit: Hi Philadelphia area friends, I will be moving Philadelphia soon and I am looking for some popular local restaurants I can turn into my staples …

Target completion for LLM:

A user likes the following recommendations: Rue De La Course in New Orleans, Swirl Wine Bar & Market in New Orleans …
The user wrote the following reviews: The wide variety of tea is impressive. On Fridays they offer free wine tastings …
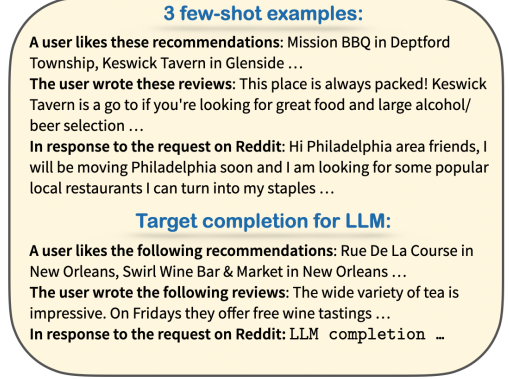**In response to the request on Reddit:** `LLM completion …`

Fig. 2. The format of the prompt used in MINT for generating synthetic narrative queries from user-item interaction with a large language model.

However, with the emergence of conversational interfaces for information access tasks, support for complex NDR tasks is likely to become necessary. In this context, recent work has noted an increase in complex and subjective natural language requests compared to more conventional search interfaces [13, 34]. Furthermore, the emergence of large language models (LLM) with strong language understanding capabilities presents the potential for fulfilling such complex requests [9, 33]. This work explores the potential for re-purposing historical user-item recommendation datasets, traditionally used for training collaborative filtering recommenders, with LLMs to support NDR.

Specifically, given a user's interactions, D_u, with items and their accompanying text documents (e.g., reviews, descriptions) $D_u = \{d_i\}_{i=1}^{N_u}$, selected from a user-item interaction dataset $\mathcal{I}$, we prompt InstructGPT, a 175B parameter LLM, to author a synthetic narrative query q_u based on D_u (Figure 2). Since we expect the query q_u to be noisy and not fully representative of all the user reviews, D_u is filtered to retain only a fraction of the reviews based on a language-model assigned likelihood of q_u given a user document, d_i. Then, a pre-trained LM based retrieval model (110M parameters) is fine-tuned for retrieval on the synthetic queries and filtered reviews.

Our approach, which we refer to as MINT[2], follows from the observation that while narrative queries and suggestions are often made in online discussion forums, and could serve as training data, the number of these posts and the diversity of domains for which they are available is significantly smaller than the size and diversity of passively gathered user-item interaction datasets. E.g. while Bogers and Koolen [5] note nearly 25,000 narrative requests for books on the LibraryThing discussion forum, a publicly available user-item interaction dataset for Goodreads contains interactions with nearly 2.2M books by 460k users [43] .

We empirically evaluate MINT in a publicly available test collection for point of interest recommendation: POINTREC [1]. To train our NDR models, we generate synthetic training data based on user-item interaction datasets from Yelp. Models (110M parameters) trained with MINT significantly outperform several baseline models and match the performance of significantly larger LLM baselines autoregressively generating recommendations. Code and synthetic datasets are available:https://github.com/iesl/narrative-driven-rec-mint/

---

[2]MINT: Data augMentation with INteraction narraTives.

2 RELATED WORK

Data Augmentation for Information Access. A line of recent work has explored using language models to generate synthetic queries for data augmentation to train models for information retrieval tasks [7, 8, 15, 23, 31]. Here, given a document collection of interest, a pre-trained language model is used to create synthetic queries for the document collection. An optional filtering step excludes noisy queries, and finally, a bi-encoder or a cross-encoder is trained for the retrieval task. While earlier work of Ma et al. [31] train a custom query generation model on web-text datasets, more recent work has leveraged large language models for zero/few-shot question generation [7, 8, 15, 23]. In generating synthetic queries, this work indicates the effectiveness of smaller parameter LLMs (up to 6B parameters) for generating synthetic queries in simpler information-retrieval tasks [7, 8, 23], and finds larger models (100B parameters and above) to be necessary for harder tasks such as argument retrieval [15, 23]. Similar to this work, we explore the generation of synthetic queries with LLMs for a retrieval task. Unlike this work, we demonstrate a data augmentation method for creating effective training data from *sets* of user documents found in recommendation datasets rather than individual documents. Other work in this space has also explored training more efficient multi-vector models from synthetic queries instead of more expensive cross-encoder models [39] and generating queries with a diverse range of intents than the ones available in implicit feedback datasets to enhance item retrievability [35].

Besides creating queries for ad-hoc retrieval tasks, concurrent work of Leszczynski et al. [25] has also explored the creation of synthetic *conversational* search datasets from music recommendation datasets with LLMs. The synthetic queries and user documents are then used to train bi-encoder retrieval models for conversational search. Our work resembles this in creating synthetic queries from *sets* of user items found in recommendation interaction datasets. However, it differs in the task of focus, creating long-form narrative queries for NDR. Finally, our work also builds on the recent perspective of Radlinski et al. [36] who make a case for natural language user profiles driving recommenders – narrative requests tie closely to natural language user profiles. Our work presents a step toward these systems.

Finally, while our work explores data augmentation from user-item interactions for a retrieval-oriented NDR task, prior work has also explored data augmentation of the user-item graph for training collaborative filtering models. This work has often explored augmentation to improve recommendation performance for minority [12, 47] or cold-start users [11, 28, 45]. And has leveraged generative models [11, 45] and text similarity models [28] for augmenting the user-item graph.

Complex Queries in Information Access. With the advent of performant models for text understanding, focus on complex and interactive information access tasks has seen a resurgence [2, 29, 32, 48]. NDR presents an example of this – NDR was first formalized in Bogers and Koolen [5] for the case of book recommendation and subsequently studied in other domains [3, 4, 6]. Bogers and Koolen [5] systematically examined narrative requests posted by users on discussion forums. They defined NDR as a task requiring item recommendation based on a long-form narrative query and prior-user item interactions. While this formulation resembles personalized search [42] and query-driven recommendation [20], the length and complexity of requests differentiate these from NDR. Other work has also demonstrated the effectiveness of re-ranking initial recommendations from collaborative filtering approaches based on the narrative query [18]. More recent work of Afzali et al. [1] formulate the NDR task without access to the prior interactions of a user while also noting the value of contextual cues contained in the narrative request. In our work, we focus on this latter formulation of NDR, given the lack of focus on effectively using the rich narrative queries in most prior work. Further, we demonstrate the usefulness of data augmentation from LLMs and user-item interaction datasets lacking narrative queries.
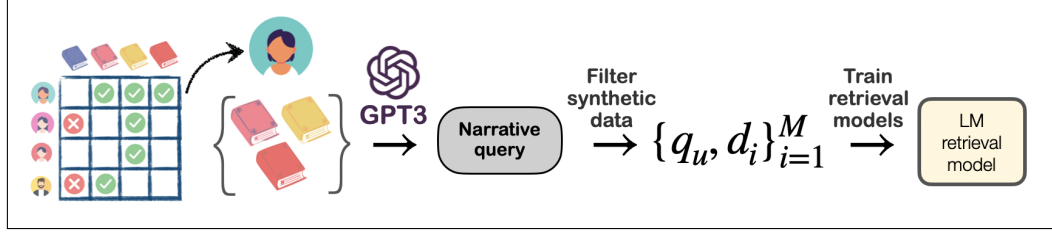
Fig. 3. MINT re-purposes readily available user-item interaction datasets commonly used to train collaborative filtering models for narrative-driven recommendation. This is done by authoring narrative queries for sets of items liked by a user with a large language model. The data is filtered with a smaller language model and retrieval models are trained on the synthetic queries and user items.

Besides this, a range of work has explored more complex, long-form, and interactive query formulations for information access; these resemble queries in NDR. Arguello et al. [2] define the tip of tongue retrieval task, a known-item search task where user queries describe the rich context of items while being unable to recall item metadata itself. Mysore et al. [32] formulate an aspect conditional query-by example task where results must match specific aspects of a long natural language query. And finally, a vibrant body of work has explored conversational critiquing of recommenders where natural language feedback helps tune the recommendations received by users [30, 44, 49].

3 METHOD

3.1 Problem Setup

In our work, we define narrative-driven recommendation (NDR) to be a ranking task, where given a narrative query q made by a user u, a ranking system f must generate a ranking R over a collection of items C. Further, we assume access to a user-item interaction dataset $\mathcal{I}$ consisting of user interactions with items $(u, \{d_i\}_{i=1}^{N_u})$. We assume the items d_i to be textual documents like reviews or item descriptions. While we don't assume there to be any overlap in the users making narrative queries or the collection of items C and the user-items interaction dataset $\mathcal{I}$, we assume them to be from the same broad domain, e.g., books, movies, points-of-interest.

3.2 Proposed Method

Our proposed method, MINT, for NDR, re-purposes a dataset of abundantly available user-item interactions, $\mathcal{I} = \{(u, \{d_i\}_{i=1}^{N_u})\}$ into training data for retrieval models by using LLMs as query generation models to author narrative queries q_u: $\mathcal{D} = \{(q_u, \{d_i\}_{i=1}^{N_u})\}$. Then, retrieval models are trained on the synthetic dataset $\mathcal{D}$ (Figure 3).

3.2.1 Narrative Queries from LLMs. To author a narrative query q_u for a user in $\mathcal{I}$, we make use of the 175B parameter InstructGPT[3] model as our query generation model QGEN. We include the text of interacted items $\{d_i\}_{i=1}^{N_u}$ in the prompt for QGEN, and instruct it to author a narrative query (Figure 2). To improve the coherence of generated queries and obtain correctly formatted outputs, we manually author narrative queries for 3 topically diverse users based on their interacted items and include it in the prompt for QGEN. The same three few shot examples are used for the whole dataset $\mathcal{I}$, and the three users were chosen from $\mathcal{I}$. Generating narrative queries based on user interactions may also be considered a form of multi-document summarization for generating a natural language user profile [36].

---

[3]https://platform.openai.com/docs/models/gpt-3, `text-davinci-003`

3.2.2 Filtering Items for Synthetic Queries. Since we expect user items to capture multiple aspects of their interests and generated queries to only capture a subset of these interests, we only retain some of the items present in $\{d_i\}_{i=1}^{N_u}$ before using it for training retrieval models. For this, we use a pre-trained language model to compute the likelihood of the query given each user item, $P_{LM}(q_u|d_i)$, and only retain the top M highly scoring item for q_u, this results in M training samples per user for our NDR retrieval models: $\{(q_u, d_i)_{i=1}^M\}$. In our experiments, we use FlanT5 with 3B parameters [14] for computing and follow Sachan et al. [40] for computing $P_{LM}(q_u|d_i)$. Note that our use of $P_{LM}(q_u|d_i)$ represents a query-likelihood model classically used for ad-hoc search and recently shown to be an effective unsupervised re-ranking method when used with large pre-trained language models [40].

3.2.3 Training Retrieval Models. We train bi-encoder and cross-encoder models for NDR on the generated synthetic dataset – commonly used models in search tasks. Bi-encoders are commonly used as scalable first-stage rankers from a large collection of items. On the other hand, cross-encoders allow a richer interaction between query and item and are used as second-stage re-ranking models. For both models, we use a pre-trained transformer language model architecture with 110M parameters, MPNet, a model similar to Bert [41]. Bi-encoder models embed the query and item independently into high dimensional vectors: $\mathbf{q}_u = \text{MPNet}(q_u)$, $\mathbf{d}_i = \text{MPNet}(d_i)$ and rank items for the user based on the minimum L2 distance between $\mathbf{q}_u$ and $\mathbf{d}_i$. Embeddings are obtained by averaging token embeddings from the final layer of MPNet, and the same model is used for both queries and items. Cross-encoder models input both the query and item and output a score to be used for ranking $s = f_{\text{CR}}([q_u; d_i])$, where f_{CR} is parameterized as $\mathbf{w}^T \text{dropout}\left(\mathbf{W}^T\text{MPNet}(\cdot)\right)$. We train our bi-encoder model with a margin ranking loss: $\mathcal{L}_{Bi} = \sum_u \sum_{i=1}^M \max[L2(\mathbf{q}_u, \mathbf{d}_i) - L2(\mathbf{q}_u, \mathbf{d}_i') + \delta, 0]$ with randomly sampled negatives d' and $\delta = 1$. Our cross-encoders are trained with a cross-entropy loss: $\mathcal{L}_{Cr} = \sum_u \sum_{i=1}^M \log(\frac{e^s}{\sum_{d'} e^{s'}})$. For training, 4 negative example items d' are randomly sampled from ranks 100-300 from our trained bi-encoder. At test time, we retrieve the top 200 items with our trained bi-encoder and re-rank them with the cross-encoder - we evaluate both these components in experiments and refer to them as BiEnc-Mint and CrEnc-Mint.

4 EXPERIMENTS AND RESULTS

Next, we evaluate Mint on a publicly available test collection for NDR and present a series of ablations.

4.1 Experimental Setup

4.1.1 Datasets. We perform evaluations on an NDR dataset for point-of-interest (POI) recommendation Pointrec [1]. Pointrec contains 112 realistic narrative queries (130 words long) obtained from discussion forums on Reddit and items pooled from baseline rankers. The items are annotated on a graded relevance scale by crowd-workers and/or discussion forum members and further validated by the dataset authors. The item collection C in Pointrec contains 700k POIs with metadata (category, city) and noisy text snippets describing the POI obtained from the Bing search engine. For test time ranking, we only rank the candidate items in the city and request category (e.g., "Restaurants") of the query available in Pointrec - this follows prior practice to exclude clearly irrelevant items [1, 26]. We use user-item interaction datasets from Yelp to generate synthetic queries for training.[4] Note also that we limit our evaluations to Pointrec since it presents the only publicly available, manually annotated, and candidate pooled test collection for NDR, to our knowledge. Other datasets for NDR use document collections that are no longer publicly accessible [24], contain sparse and noisy relevance judgments due to them being determined with automatic rules applied to discussion

---

[4]https://www.yelp.com/dataset

threads [18, 24], lack pooling to gather candidates for judging relevance [18, 24], or lack realistic narrative queries [21]. We leave the development of more robust test collections and evaluation methods for NDR to future work.

4.1.2 Implementation Details. Next, we describe important details for Mint and leave finer details of the model and training to our code release. To sample user interactions for generating synthetic queries from the Yelp dataset, we exclude POIs and users with fewer than ten reviews to ensure that users were regular users of the site with well represented interests. This follows common prior practice in preparing user-item interaction datasets for use [27]. Then we retain users who deliver an average rating greater than 3/5 and with 10-30 above-average reviews. This desirably biases our data to users who commonly describe their likings (rather than dislikes). It also retains the users whose interests are summarizable by QGen. In the Yelp dataset, this results in 45,193 retained users. Now, 10,000 randomly selected users are chosen for generating synthetic narrative queries. For these users, a single randomly selected sentence from 10 of their reviews is included in the prompt (Figure 2) to QGen, i.e., $N_u = 10$. After generating synthetic queries, some items are filtered out (§3.2.2). Here, we exclude 40% of the items for a user. This results in about 60,000 training samples for training BiEnc-Mint and CrEnc-Mint. These decisions were made manually by examining the resulting datasets and the cost of authoring queries. The expense of generating q_u was about USD 230.

4.1.3 Baselines. We compare BiEnc-Mint and CrEnc-Mint models against several standard and performant retrieval model baselines. These span zero-shot/unsupervised rankers, supervised bi-encoders, unsupervised cross-encoders, and LLM baselines. BM25: A standard unsupervised sparse retrieval baseline based on term overlap between query and document, with strong generalization performance across tasks and domains [38]. Contriver: A BERT-base bi-encoder model pre-trained for zero-shot retrieval with weakly supervised query-document pairs [22]. MPNet-1B: A strong Sentence-Bert bi-encoder model initialized with MPNet-base and trained on 1 billion supervised query-document pairs aggregated from numerous domains [37]. BERT-MSM: A BERT-base bi-encoder fine-tuned on supervised question-passage pairs from MSMarco. UPR: A two-stage approach that retrieves items with a Contriver bi-encoder and re-ranks the top 200 items with a query-likelihood model using a FlanT5 model with 3B parameters [14, 40]. This may be seen as an unsupervised "cross-encoder" model. Grounded LLM: A recently proposed two-stage approach which autoregressively generates ten pseudo-relevant items using an LLM (175B InstructGPT) prompted with the narrative query and generates recommendations grounded in C by retrieving the nearest neighbors for each generated item using a bi-encoder [19]. We include one few-shot example of a narrative query and recommended items in the prompt to the LLM. We run this baseline three times and report average performance across runs. We report NDCG at 5 and 10, MAP, MRR, and Recall at 100 and 200. Finally, our reported results should be considered lower bounds on realistic performance due to the unjudged documents (about 70% at $k = 10$) in our test collections [10].

4.2 Results

Table 1 presents the performance of the proposed method compared against baselines. Here, bold numbers indicate the best-performing model, and superscripts indicate statistical significance computed with two-sided t-tests at $p < 0.05$.

Here, we first note the performance of baseline approaches. We see BM25 outperformed by Contriver, a transformer bi-encoder model trained for zero-shot retrieval; this mirrors prior work [22]. Next, we see supervised bi-encoder models trained on similar passage (MPNet-1B) and question-answer (BERT-MSM) pairs outperform a weakly supervised model (Contriver) by smaller margins. Finally, the Grounded LLM outperforms all bi-encoder baselines, indicating strong few-shot generalization and mirroring prior results [19]. Examining the Mint models, we first note that the BiEnc-Mint sees statistically significant improvement compared to BM25 and outperforms the best bi-encoder baselines by

Table 1. Performance of the proposed method, Mint, for point-of-interest recommendation on Pointrec. The superscripts denote statistically significant improvements compared to specific baseline models.

| Model | Parameters | POINTREC | | | | | |
		NDCG@5	NDCG@10	MAP	MRR	Recall@100	Recall@200
[1]BM25	-	0.2682	0.2464	0.1182	0.2685	0.4194	0.5429
[2]Contriver	110M	0.2924	0.2776	0.1660	0.3355	0.4455	0.5552
[3]MPNet-1B	110M	0.3038	0.2842	0.1621	0.3566	0.4439	0.5657
[4]BERT-MSM	110M	0.3117	0.2886	0.1528	0.3320	0.4679	0.5816
[5]Grounded LLM	175B+110M	0.3558	0.3251	0.1808	0.3861	0.4797	0.5797
[6]UPR	110M+3B	0.3586	0.3242	0.1712	0.4013	0.4489	0.5552
BiEnc-Mint	110M	0.3489[1]	0.3263[1]	0.1890[1]	0.3982[1]	0.4914[1]	**0.6221**
CrEnc-Mint	2×110M	**0.3725**[12]	**0.3489**[12]	**0.2192**[14]	**0.4317**[1]	**0.5448**[123]	**0.6221**

11-13% on precision measures and 5-7% on recall measures. Specifically, we see a model trained for question-answering (BERT-MSM) underperform BiEnc-Mint, indicating the challenge of the NDR task. Further, BiEnc-Mint, trained on 5 orders of magnitude lesser data than MPNet-1B, sees improved performance – indicating the quality of data obtained from Mint. Furthermore, BiEnc-Mint also performs at par with a 175B LLM while offering the inference efficiency of a small-parameter bi-encoder. Next, we see CrEnc-Mint outperform the baseline bi-encoders, BiEnc-Mint, UPR, and Grounded LLM by 4-21% on precision measures and 7-13% on recall measures – demonstrating the value of augmentation with Mint for training NDR models.

4.3 Ablations

In Table 2, we ablate various design choices in Mint. Different choices result in different training sets for the BiEnc and CrEnc models. Also, note that in reporting ablation performance for CrEnc, we still use the performant BiEnc-Mint model for obtaining negative examples for training and first-stage ranking. Without high-quality negative examples, we found CrEnc to result in much poorer performance.

No item filtering. Since synthetic queries are unlikely to represent all the items of a user, Mint excludes user items $\{d_i\}_{i=1}^{N_u}$ which have a low likelihood of being generated from the document (§3.2.2). Without this step, we expect the training set for training retrieval models to be larger and noisier. In Table 2, we see that excluding this step leads to a lower performance for BiEnc and CrEnc, indicating that the quality of data obtained is important for performance.

6B LLM for QGen. Mint relies on using an expensive 175B parameter InstructGPT model for QGen. Here, we investigate the efficacy for generating q_u for $\{d_i\}_{i=1}^{N_u}$ with a 6B parameter InstructGPT model (text-curie-001). We use an identical setup to the 175B LLM for this. In Table 2, we see that training on the synthetic narrative queries of the smaller LLM results in worse models – often underperforming the baselines in Table 1. This indicates the inability of a smaller model to generate complex narrative queries while conditioning on a set of user items. This necessity of a larger LLM for generating queries in complex retrieval tasks has been observed in prior work [15, 23].

6B LLM for Item Queries. We find a smaller 6B LLM to result in poor quality data when used to generate narrative queries conditioned on $\{d_i\}_{i=1}^{N_u}$. Here we simplify the text generation task – using a 6B LLM to generate queries for individual items d_i. This experiment also mirrors the setup for generating synthetic queries for search tasks [7, 15]. Here, we use 3-few shot examples and sample one item per user for generating q_u. Given the lower cost of using a smaller LLM, we use all 45,193 users in our Yelp dataset rather than a smaller random sample. From Table 2, we see

Table 2. Mint ablated for different design choices on Pointrec.

Ablation	Pointrec					
	NDCG@5	NDCG@10	MAP	MRR	Recall@100	Recall@200
BiEnc-Mint	0.3489	0.3263	0.1890	0.3982	0.5263	0.6221
− No item filtering	0.2949	0.2766	0.1634	0.3505	0.4979	0.5951
− 6B LLM for QGen	0.2336	0.2293	0.1125	0.2287	0.426	0.5435
− 6B LLM for Item Queries	0.3012	0.2875	0.1721	0.3384	0.4800	0.5909
CrEnc-Mint	0.3725	0.3489	0.2192	0.4317	0.5448	0.6221
− No item filtering	0.3570	0.3379	0.2071	0.4063	0.5366	0.6221
− 6B LLM for QGen	0.2618	0.2421	0.1341	0.3118	0.4841	0.6221
− 6B LLM for Item Queries	0.3792	0.3451	0.2128	0.4098	0.5546	0.6221

that this results in higher quality queries than using smaller LLMs for generating narrative queries from $\{d_i\}_{i=1}^{N_u}$. The resulting BiEnc model underperforms the BiEnc-Mint, indicating the value of generating complex queries conditioned on multiple items as in Mint for NDR. We see that CrEnc approaches the performance of CrEnc-Mint– note, however, that this approach uses the performant BiEnc-Mint for sampling negatives and first stage ranking. We leave further exploration of using small parameter LLMs for data augmentation for NDR models to future work.

5 CONCLUSIONS

In this paper, we present Mint, a data augmentation method for the narrative-driven recommendation (NDR) task. Mint re-purposes historical user-item interaction datasets for NDR by using a 175B parameter large language model to author long-form narrative queries while conditioning on the text of items liked by users. We evaluate bi-encoder and cross-encoder models trained on data from Mint on the publicly available Pointrec test collection for narrative-driven point of interest recommendation. We demonstrate that the resulting models outperform several strong baselines and ablated models and match or outperform a 175B LLM directly used for NDR in a 1-shot setup.

However, Mint also presents some limitations. Given our use of historical interaction datasets for generating synthetic training data and the prevalence of popular interests in these datasets longer, tailed interests are unlikely to be present in the generated synthetic datasets. In turn, causing retrieval models to likely see poorer performance on these requests. Our use of LLMs to generate synthetic queries also causes the queries to be repetitive in structure, likely causing novel longer-tail queries to be poorly served. These limitations may be addressed in future work.

Besides this, other avenues also present rich future work. While Mint leverages a 175B LLM for generating synthetic queries, smaller parameter LLMs may be explored for this purpose - perhaps by training dedicated QGen models. Mint may also be expanded to explore more active strategies for sampling items and users for whom narrative queries are authored - this may allow more efficient use of large parameter LLMs while ensuring higher quality training datasets. Next, the generation of synthetic queries from *sets* of documents may be explored for a broader range of retrieval tasks beyond NDR given its promise to generate larger training sets – a currently underexplored direction. Finally, given the lack of larger-scale test collections for NDR and the effectiveness of LLMs for authoring narrative queries from user-item interaction, fruitful future work may also explore the creation of larger-scale datasets in a mixed-initiative setup to robustly evaluate models for NDR tasks.

ACKNOWLEDGMENTS

REFERENCES

[1] Jafar Afzali, Aleksander Mark Drzewiecki, and Krisztian Balog. 2021. POINTREC: A Test Collection for Narrative-Driven Point of Interest Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2478–2484. https://doi.org/10.1145/3404835.3463243

[2] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the 6th international ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM. https://dlnext.acm.org/doi/10.1145/3406522.3446021

[3] Toine Bogers, Maria Gäde, Marijn Koolen, Vivien Petras, and Mette Skov. 2018. "What was this Movie About this Chick?" A Comparative Study of Relevance Aspects in Book and Movie Discovery. In *Transforming Digital Worlds: 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings 13*. Springer, 323–334.

[4] Toine Bogers, Maria Gäde, Marijn Koolen, Vivien Petras, and Mette Skov. 2019. "Looking for an amazing game I can relax and sink hours into...": A Study of Relevance Aspects in Video Game Discovery. In *Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings 14*. Springer, 503–515.

[5] Toine Bogers and Marijn Koolen. 2017. Defining and Supporting Narrative-Driven Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 238–242. https://doi.org/10.1145/3109859.3109893

[6] Toine Bogers and Marijn Koolen. 2018. "I'm looking for something like…": Combining Narratives and Example Items for Narrative-driven Book Recommendation. In *Knowledge-aware and Conversational Recommender Systems Workshop*. CEUR Workshop Proceedings.

[7] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2387–2392. https://doi.org/10.1145/3477495.3531863

[8] Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2023. InPars-Light: Cost-Effective Unsupervised Training of Efficient Rankers. arXiv:2301.02998

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[10] Chris Buckley and Ellen M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, United Kingdom) *(SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/1008992.1009000

[11] Dong-Kyu Chae, Jihoo Kim, Duen Horng Chau, and Sang-Wook Kim. 2020. AR-CF: Augmenting Virtual Users and Items in Collaborative Filtering for Addressing Cold-Start Problems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1251–1260. https://doi.org/10.1145/3397271.3401038

[12] Lei Chen, Le Wu, Kun Zhang, Richang Hong, Defu Lian, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2023. Improving Recommendation Fairness via Data Augmentation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 1012–1020. https://doi.org/10.1145/3543507.3583341

[13] Li Chen, Zhirun Zhang, Xinzhi Zhang, and Lehong Zhao. 2022. A Pilot Study for Understanding Users' Attitudes Towards a Conversational Agent for News Recommendation. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 36, 6 pages. https://doi.org/10.1145/3543829.3544530

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[15] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=gmL46YMpu2J

[16] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/1242572.1242610

[17] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. 2010. The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 293–296. https://doi.org/10.1145/1864708.1864770

[18] Lukas Eberhard, Simon Walk, Lisa Posch, and Denis Helic. 2019. Evaluating Narrative-Driven Movie Recommendations on Reddit. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3301275.3302287

[19] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. *arXiv preprint arXiv:2212.10496* (2022).

[20] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2013. Query-Driven Context Aware Recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 9–16. https://doi.org/10.1145/2507157.2507187

[21] Seyyed Hadi Hashemi, Jaap Kamps, Julia Kiseleva, Charles LA Clarke, and Ellen M Voorhees. 2016. Overview of the TREC 2016 Contextual Suggestion Track.. In *TREC*.

[22] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=jKN1pXi7b0

[23] Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval. arXiv:2301.01820

[24] Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh. 2016. Overview of the CLEF 2016 Social Book Search Lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 351–370.

[25] Megan Leszczynski, Ravi Ganti, Shu Zhang, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. Generating Synthetic Data for Conversational Music Recommendation Using Random Walks and Language Models. arXiv:2301.11489

[26] Xin Liu, Yong Liu, Karl Aberer, and Chunyan Miao. 2013. Personalized Point-of-Interest Recommendation by Mining Users' Preference Transition. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 733–738. https://doi.org/10.1145/2505515.2505639

[27] Yiding Liu, Tuan-Anh Nguyen Pham, Gao Cong, and Quan Yuan. 2017. An Experimental Evaluation of Point-of-Interest Recommendation in Location-Based Social Networks. *Proc. VLDB Endow.* 10, 10 (jun 2017), 1010–1021. https://doi.org/10.14778/3115404.3115407

[28] Federico López, Martin Scholz, Jessica Yung, Marie Pellat, Michael Strube, and Lucas Dixon. 2021. Augmenting the user-item graph with textual similarity models. *arXiv preprint arXiv:2109.09358* (2021).

[29] Xing Han Lu, Siva Reddy, and Harm de Vries. 2023. The StatCan Dialogue Dataset: Retrieving Data Tables through Conversations with Genuine Intents. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 2799–2829. https://aclanthology.org/2023.eacl-main.206

[30] Kai Luo, Scott Sanner, Ga Wu, Hanze Li, and Hojin Yang. 2020. Latent Linear Critiquing for Conversational Recommender Systems. In *The Web Conference*.

[31] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1075–1088. https://doi.org/10.18653/v1/2021.eacl-main.92

[32] Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - A Test Collection of Computer Science Research Articles for Faceted Query by Example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://doi.org/10.48550/arXiv.2103.12906

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[34] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Alfred Sliwa, Ahmet Aker, and Norbert Fuhr. 2021. Starting Conversations with Search Engines - Interfaces That Elicit Natural Language Queries. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 261–265. https://doi.org/10.1145/3406522.3446035

[35] Gustavo Penha, Enrico Palumbo, Maryam Aziz, Alice Wang, and Hugues Bouchard. 2023. Improving Content Retrievability in Search with Controllable Query Generation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3182–3192. https://doi.org/10.1145/3543507.3583261

[36] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2863–2874. https://doi.org/10.1145/3477495.3531873

[37] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[38] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[39] Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Arafat Sultan, and Christopher Potts. 2023. UDAPDR: Unsupervised Domain Adaptation via LLM Prompting and Distillation of Rerankers. arXiv:2303.00807 [cs.IR]

[40] Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving Passage Retrieval with Zero-Shot Question Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3781–3797. https://aclanthology.org/2022.emnlp-main.249

[41] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, Vol. 33. https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

[42] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing Search via Automated Analysis of Interests and Activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 449–456. https://doi.org/10.1145/1076034.1076111

[43] Mengting Wan and Julian McAuley. 2018. Item Recommendation on Monotonic Behavior Chains. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 86–94. https://doi.org/10.1145/3240323.3240369

[44] Haonan Wang, Chang Zhou, Carl Yang, Hongxia Yang, and Jingrui He. 2021. Controllable Gradient Item Retrieval. In *Web Conference*.

[45] Qinyong Wang, Hongzhi Yin, Hao Wang, Quoc Viet Hung Nguyen, Zi Huang, and Lizhen Cui. 2019. Enhancing Collaborative Filtering with Generative Augmentation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 548–556. https://doi.org/10.1145/3292500.3330873

[46] Jiajing Xu, Andrew Zhai, and Charles Rosenberg. 2022. Rethinking Personalized Ranking at Pinterest: An End-to-End Approach. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 502–505. https://doi.org/10.1145/3523227.3547394

[47] Yuxin Ying, Fuzhen Zhuang, Yongchun Zhu, Deqing Wang, and Hongwei Zheng. 2023. CAMUS: Attribute-Aware Counterfactual Augmentation for Minority Users in Recommendation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 1396–1404. https://doi.org/10.1145/3543507.3583538

[48] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808* (2022).

[49] Jie Zou, Yifan Chen, and Evangelos Kanoulas. 2020. Towards Question-Based Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 881–890. https://doi.org/10.1145/3397271.3401180