LLM-Personalized Retrieval: Query Rewriting and Negatives for Session Search

B10705005 陳思如 B10705009 邱一新

Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

Introduction

- Tradition search systems rely on static data
- LLMs offer ability to enhance personalization
  - understanding user profile and session context
  - generate relevant and dynamic queries
  - without human interference cost
- Negative data give model more information to distinguish the differences
- Goal
  - More personalized and up-to-date search context
  - Explore rewrite prompt strategy
  - Improve the relevance and accuracy of the personalized search results

Related Work - Query Rewrite

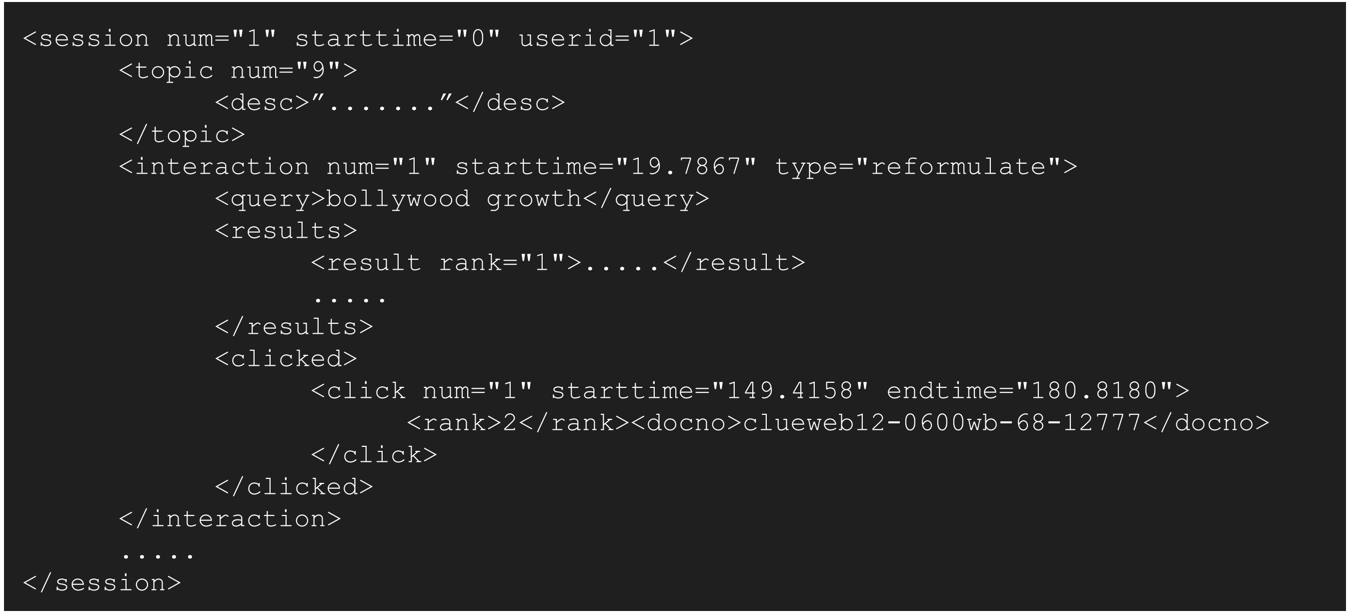
- *Query Rewriting via Large Language Models*
- *R-Bot: An LLM-based Query Rewrite System*
- *Don't Retrieve, Generate: Prompting LLMs for Synthetic Training Data in Dense Retrieval*

Related Work - Hard Negative

- *Optimizing Dense Retrieval Model Training with Hard Negatives*
- *Passage-based BM25 Hard Negatives: A Simple and Effective Negative Sampling Strategy For Dense Retrieval*
- *TriSampler: A Better Negative Sampling Principle for Dense Retrieval*

Methodology - Session Data Construction

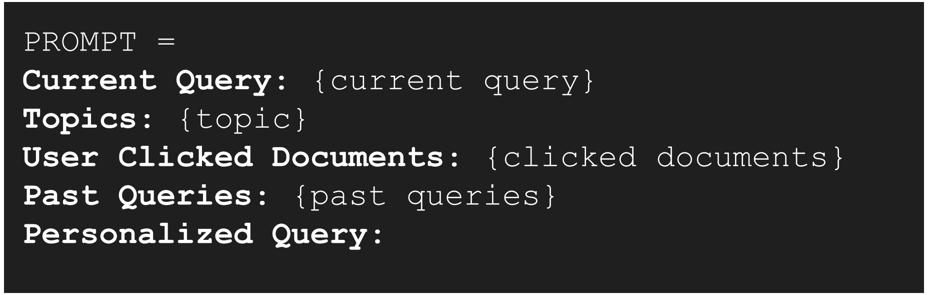
- Dataset: TREC 2014 Session Track
- Total: 1041 samples (train: 936 / test: 105)
- Take current interaction query and first clicked document as training pairs



```
<session num="1" starttime="0" userid="1">
      <topic num="9">
            <desc>”.......”</desc>
      </topic>
      <interaction num="1" starttime="19.7867" type="reformulate">
            <query>bollywood growth</query>
            <results>
                  <result rank="1">.....</result>
                  .....
            </results>
            <clicked>
                  <click num="1" starttime="149.4158" endtime="180.8180">
                        <rank>2</rank><docno>clueweb12-0600wb-68-12777</docno>
                  </click>
            </clicked>
      </interaction>
      .....
</session>
```

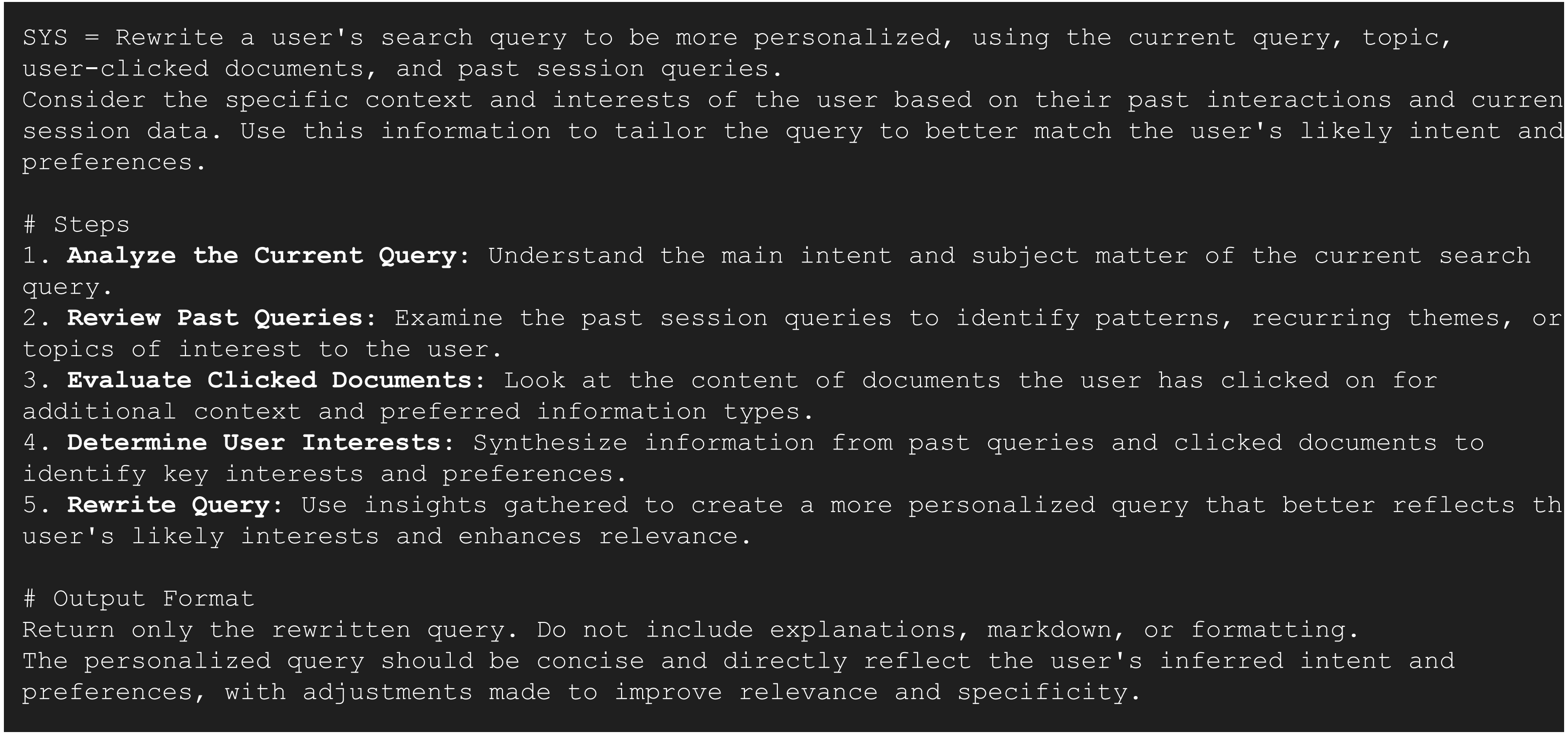
Methodology - Rewrite Strategy

- For every training pairs, collect
    - user id
    - current topic
    - prior interaction queries in the section
    - prior clicked documents in the section



```
PROMPT =
Current Query: {current query}
Topics: {topic}
User Clicked Documents: {clicked documents}
Past Queries: {past queries}
Personalized Query:
```

Methodology - Rewrite Prompt #1 (Long Query)

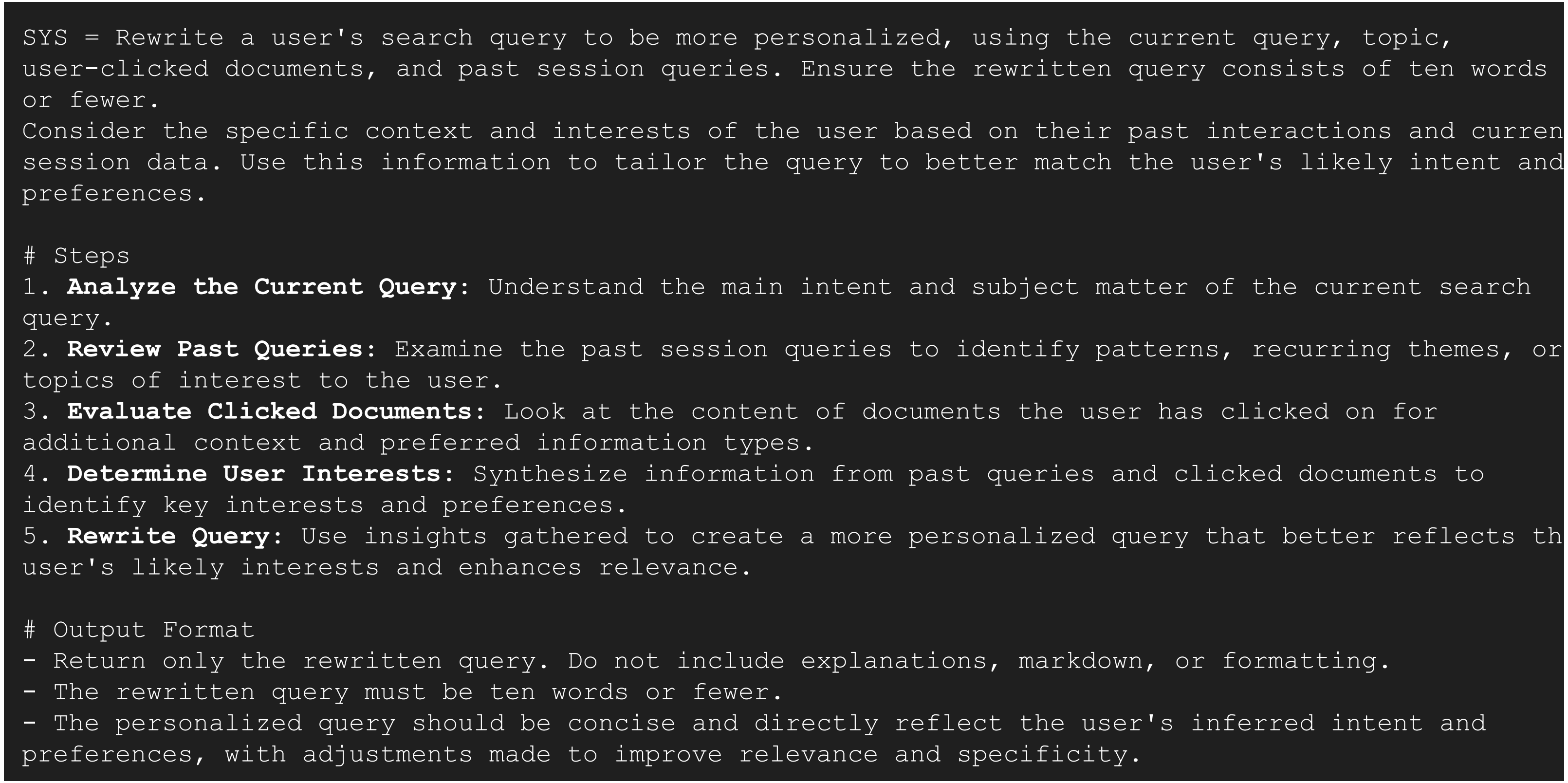


```
SYS = Rewrite a user's search query to be more personalized, using the current query, topic,
user-clicked documents, and past session queries.
Consider the specific context and interests of the user based on their past interactions and curren
session data. Use this information to tailor the query to better match the user's likely intent and
preferences.

# Steps
1. Analyze the Current Query: Understand the main intent and subject matter of the current search
query.
2. Review Past Queries: Examine the past session queries to identify patterns, recurring themes, or
topics of interest to the user.
3. Evaluate Clicked Documents: Look at the content of documents the user has clicked on for
additional context and preferred information types.
4. Determine User Interests: Synthesize information from past queries and clicked documents to
identify key interests and preferences.
5. Rewrite Query: Use insights gathered to create a more personalized query that better reflects th
user's likely interests and enhances relevance.

# Output Format
Return only the rewritten query. Do not include explanations, markdown, or formatting.
The personalized query should be concise and directly reflect the user's inferred intent and
preferences, with adjustments made to improve relevance and specificity.
```

Methodology - Rewrite Prompt #1 (Long Query)

```
Input:
{
     "query": "seattle vacation",
     "positive_doc": "clueweb12-0112wb-18-23823",
     "session_queries": ["hawaii vacation"],
     "topic_desc": "Suppose you'd like to take a week-long vacation.  Research some
     possible destinations.  How do they compare in terms of cost (travel + room and
     board + entertainment), value (things to do), and feasibility (ease of getting
     there)?"
}

Output: "Top destinations for a week-long vacation in the US: compare costs, values,
and feasibility of Hawaii, Florida, and California. Get exclusive deals and discounts
on Hawaii vacation packages, including all-inclusive options and luxury rentals.
Consider visiting Disney World in Florida or exploring the natural beauty of
California's coastline and national parks. Find the best value for your money and
make the most of your vacation time."
```

Methodology - Rewrite Prompt #2 (Short Query)



```
SYS = Rewrite a user's search query to be more personalized, using the current query, topic,
user-clicked documents, and past session queries. Ensure the rewritten query consists of ten words
or fewer.
Consider the specific context and interests of the user based on their past interactions and curren
session data. Use this information to tailor the query to better match the user's likely intent and
preferences.

# Steps
1. Analyze the Current Query: Understand the main intent and subject matter of the current search
query.
2. Review Past Queries: Examine the past session queries to identify patterns, recurring themes, or
topics of interest to the user.
3. Evaluate Clicked Documents: Look at the content of documents the user has clicked on for
additional context and preferred information types.
4. Determine User Interests: Synthesize information from past queries and clicked documents to
identify key interests and preferences.
5. Rewrite Query: Use insights gathered to create a more personalized query that better reflects th
user's likely interests and enhances relevance.

# Output Format
- Return only the rewritten query. Do not include explanations, markdown, or formatting.
- The rewritten query must be ten words or fewer.
- The personalized query should be concise and directly reflect the user's inferred intent and
preferences, with adjustments made to improve relevance and specificity.
```

Methodology - Rewrite Prompt #2 (Short Query)

```
Input:
{
      "query": "seattle vacation",
      "positive_doc": "clueweb12-0112wb-18-23823",
      "session_queries": ["hawaii vacation"],
      "topic_desc": "Suppose you'd like to take a week-long vacation.  Research some
      possible destinations.  How do they compare in terms of cost (travel + room and
      board + entertainment), value (things to do), and feasibility (ease of getting
      there)?"
}

Output: "Best vacation spots for a week-long trip: Hawaii vs. Florida vs. Canada.
Compare costs, value, and ease of travel."
```

Methodology - Rewrite Prompt #3 (keyword only)

```
SYS = Extract relevant keywords from the user's search context (topic, clicked documents, and past
queries).

# Rules
- Extract 2-3 most relevant keywords
- Keywords should be single words
- Keywords should reflect user interests and search intent
- Return keywords separated by comma

# Output
Only return the keywords separated by comma. No explanations.


**Instructions**:
- Extract 2-3 most relevant keywords from the context
- Return only the keywords separated by comma
**Keywords**:"""
```

Methodology - Rewrite Prompt #3 (keyword only)

```
Input:
{
     "query": "seattle vacation",
     "positive_doc": "clueweb12-0112wb-18-23823",
     "session_queries": ["hawaii vacation"],
     "topic_desc": "Suppose you'd like to take a week-long vacation.  Research some
     possible destinations.  How do they compare in terms of cost (travel + room and
     board + entertainment), value (things to do), and feasibility (ease of getting
     there)?"
}

Output: "seattle vacation hawaii travel"
```

Methodology - Retrieval Model Training

- BiEncoder Model:
  - sentence-transformers/msmarco-MiniLM-L-6-v3
  - BAAI/bge-base-en-v1.5
- Cosine-Similarity between query and document embeddings
- Loss:
  - Cross Entropy Loss: (query, clicked document)
  - InfoNCE Loss: (query, clicked document, negative document)

Methodology - Negative Data Selection

- Static Negative Data
  - BM25 ranked retrieved document
  - The document with highest score but not the positive document
- Dynamic Negative Data
  - Begin with BM25 retrieved negative document for warmup epochs
  - Every refresh epochs, using the trained BiEncoder to retrieve top similar documents
  - Update the training data with new negative document to increase training diversity

Experiments - Evaluation Metric

- Hit@k

  Check ground-truth document appears in retrieved result

- MRR@k

  Evaluate how early did the ground truth appears in the retrieved result list

- NDCG@k

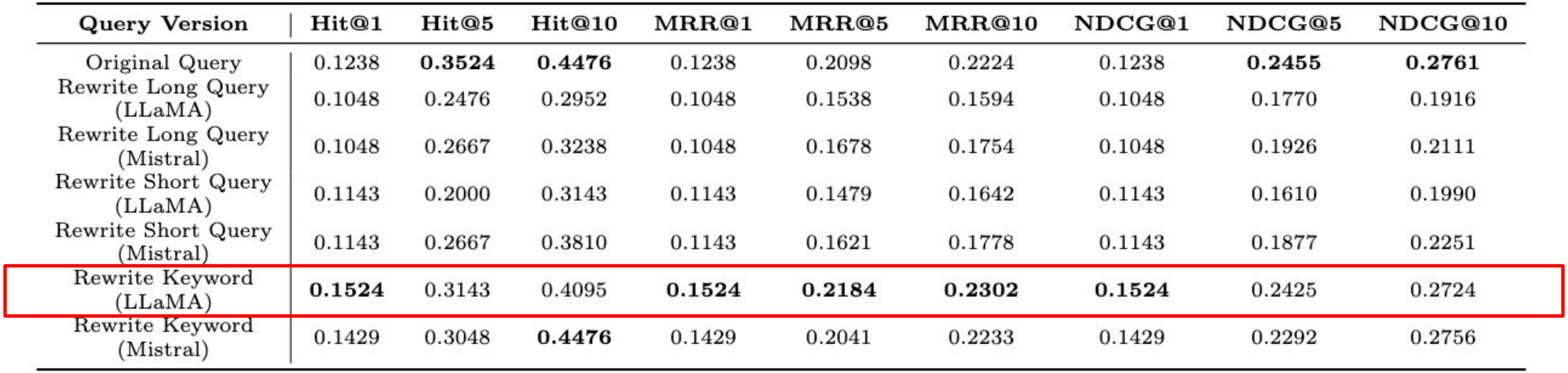
  Score ground truth document as 1 and other documents as 0.

- k = 1, 5, 10

Experiments - BM25 for Prompt Selection

Baseline method to evaluate the performance of different rewrite prompt

=> **LLaMA rewrite keyword prompt** performs the best

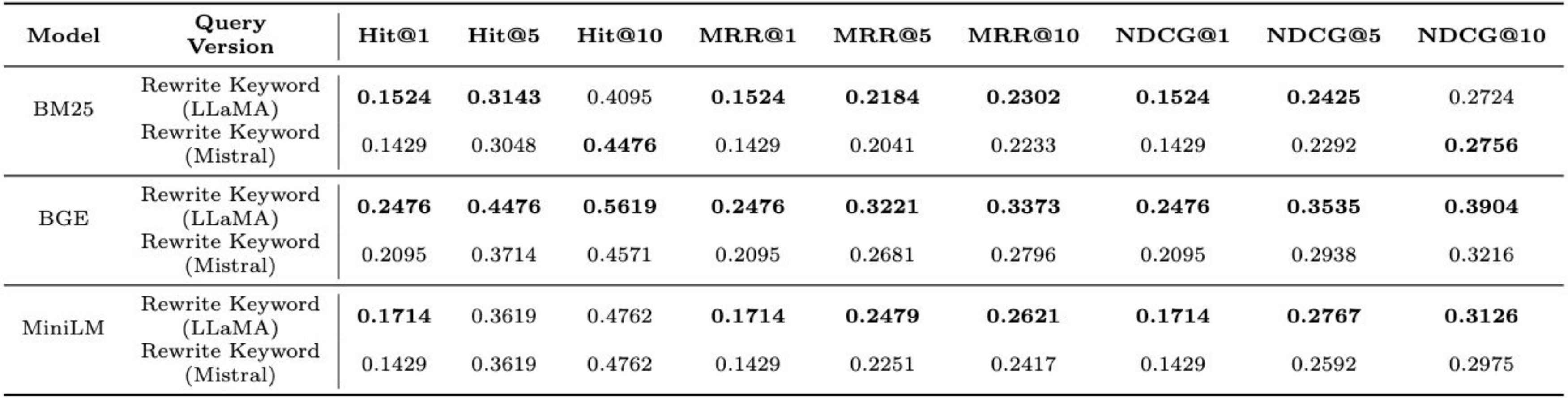


Query Version	Hit@1	Hit@5	Hit@10	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10
Original Query	0.1238	**0.3524**	**0.4476**	0.1238	0.2098	0.2224	0.1238	**0.2455**	**0.2761**
Rewrite Long Query (LLaMA)	0.1048	0.2476	0.2952	0.1048	0.1538	0.1594	0.1048	0.1770	0.1916
Rewrite Long Query (Mistral)	0.1048	0.2667	0.3238	0.1048	0.1678	0.1754	0.1048	0.1926	0.2111
Rewrite Short Query (LLaMA)	0.1143	0.2000	0.3143	0.1143	0.1479	0.1642	0.1143	0.1610	0.1990
Rewrite Short Query (Mistral)	0.1143	0.2667	0.3810	0.1143	0.1621	0.1778	0.1143	0.1877	0.2251
Rewrite Keyword (LLaMA)	**0.1524**	0.3143	0.4095	**0.1524**	**0.2184**	**0.2302**	**0.1524**	0.2425	0.2724
Rewrite Keyword (Mistral)	0.1429	0.3048	**0.4476**	0.1429	0.2041	0.2233	0.1429	0.2292	0.2756

Table 1: BM25 Performance

Experiments - Different LLM

- mistralai/Mistral-7B-Instruct-v0.1
- **meta-llama/Llama-2-7b-chat-hf**

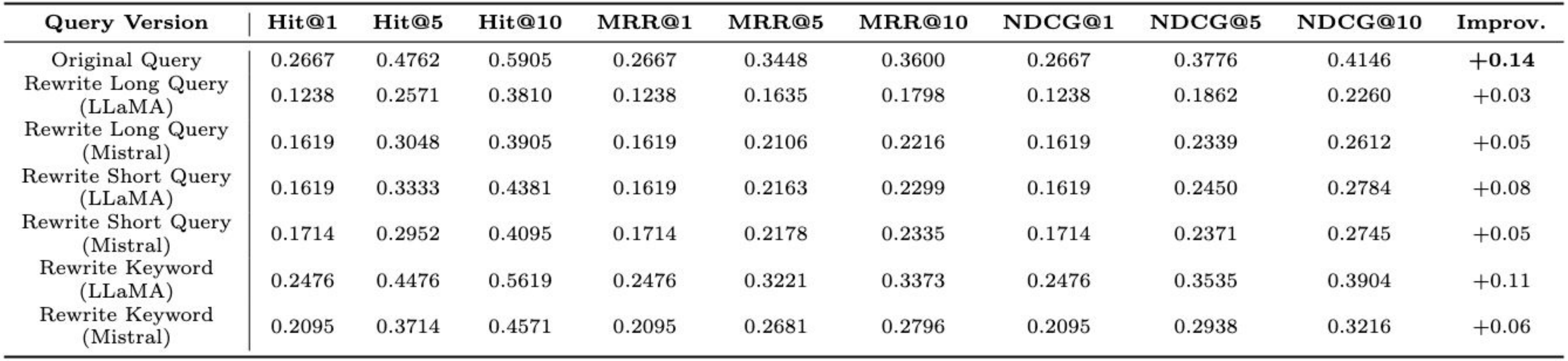


Model	Query Version	Hit@1	Hit@5	Hit@10	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10
BM25	Rewrite Keyword (LLaMA)	**0.1524**	**0.3143**	0.4095	**0.1524**	**0.2184**	**0.2302**	**0.1524**	**0.2425**	0.2724
	Rewrite Keyword (Mistral)	0.1429	0.3048	**0.4476**	0.1429	0.2041	0.2233	0.1429	0.2292	**0.2756**
BGE	Rewrite Keyword (LLaMA)	**0.2476**	**0.4476**	**0.5619**	**0.2476**	**0.3221**	**0.3373**	**0.2476**	**0.3535**	**0.3904**
	Rewrite Keyword (Mistral)	0.2095	0.3714	0.4571	0.2095	0.2681	0.2796	0.2095	0.2938	0.3216
MiniLM	Rewrite Keyword (LLaMA)	**0.1714**	0.3619	0.4762	**0.1714**	**0.2479**	**0.2621**	**0.1714**	**0.2767**	**0.3126**
	Rewrite Keyword (Mistral)	0.1429	0.3619	0.4762	0.1429	0.2251	0.2417	0.1429	0.2592	0.2975

Table 4: LLM Rewrite Performance on LLaMA Rewrite Keyword

Experiments - BiEncoder Model BGE

- BAAI/bge-base-en-v1.5
- Improves in all query version compared to BM25 baseline

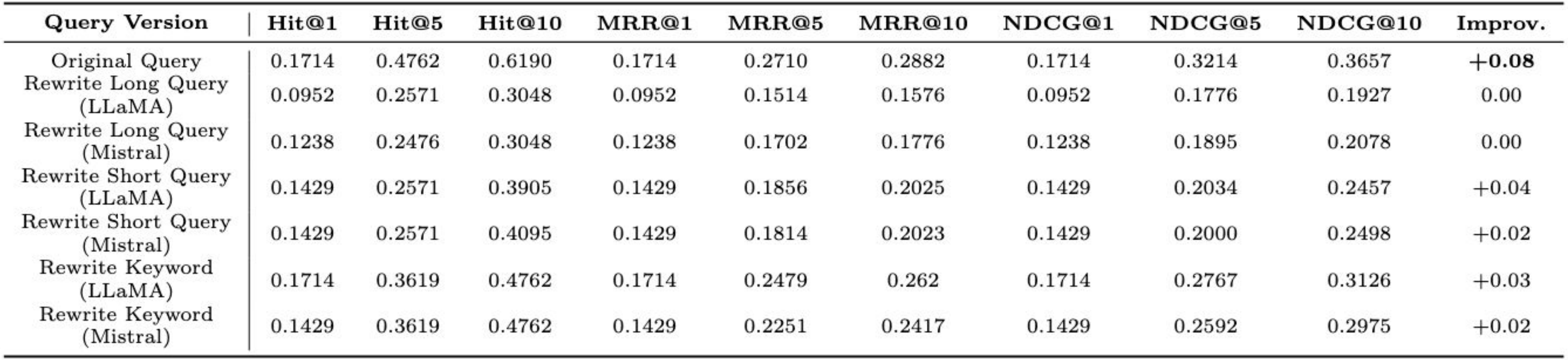


Query Version	Hit@1	Hit@5	Hit@10	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10	Improv.
Original Query	0.2667	0.4762	0.5905	0.2667	0.3448	0.3600	0.2667	0.3776	0.4146	**+0.14**
Rewrite Long Query (LLaMA)	0.1238	0.2571	0.3810	0.1238	0.1635	0.1798	0.1238	0.1862	0.2260	+0.03
Rewrite Long Query (Mistral)	0.1619	0.3048	0.3905	0.1619	0.2106	0.2216	0.1619	0.2339	0.2612	+0.05
Rewrite Short Query (LLaMA)	0.1619	0.3333	0.4381	0.1619	0.2163	0.2299	0.1619	0.2450	0.2784	+0.08
Rewrite Short Query (Mistral)	0.1714	0.2952	0.4095	0.1714	0.2178	0.2335	0.1714	0.2371	0.2745	+0.05
Rewrite Keyword (LLaMA)	0.2476	0.4476	0.5619	0.2476	0.3221	0.3373	0.2476	0.3535	0.3904	+0.11
Rewrite Keyword (Mistral)	0.2095	0.3714	0.4571	0.2095	0.2681	0.2796	0.2095	0.2938	0.3216	+0.06

Table 2: BGE BiEncoder Performance

Experiments - BiEncoder Model MiniLM

- sentence-transformers/msmarco-MiniLM-L-6-v3
- Improves in mostly query versions except Long Query rewrite version compared to BM25 baseline

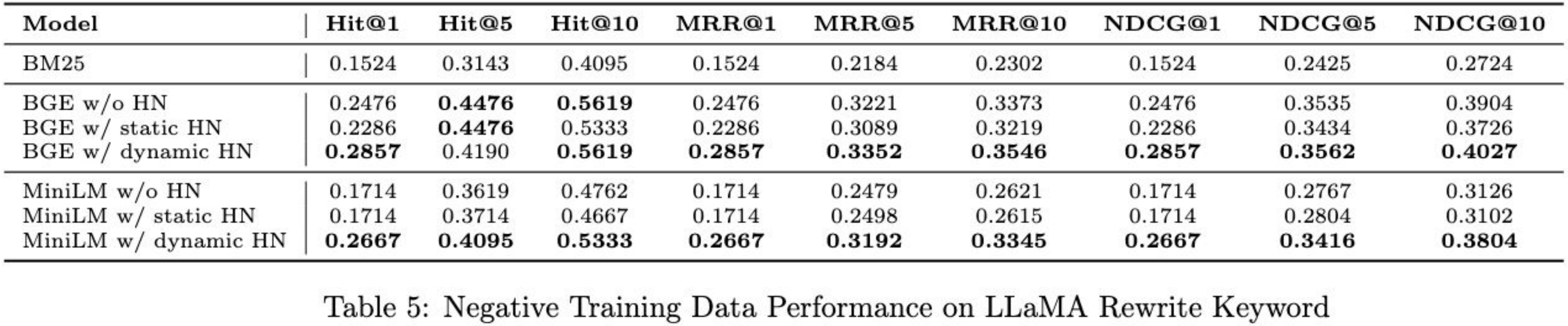


Query Version	Hit@1	Hit@5	Hit@10	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10	Improv.
Original Query	0.1714	0.4762	0.6190	0.1714	0.2710	0.2882	0.1714	0.3214	0.3657	**+0.08**
Rewrite Long Query (LLaMA)	0.0952	0.2571	0.3048	0.0952	0.1514	0.1576	0.0952	0.1776	0.1927	0.00
Rewrite Long Query (Mistral)	0.1238	0.2476	0.3048	0.1238	0.1702	0.1776	0.1238	0.1895	0.2078	0.00
Rewrite Short Query (LLaMA)	0.1429	0.2571	0.3905	0.1429	0.1856	0.2025	0.1429	0.2034	0.2457	+0.04
Rewrite Short Query (Mistral)	0.1429	0.2571	0.4095	0.1429	0.1814	0.2023	0.1429	0.2000	0.2498	+0.02
Rewrite Keyword (LLaMA)	0.1714	0.3619	0.4762	0.1714	0.2479	0.262	0.1714	0.2767	0.3126	+0.03
Rewrite Keyword (Mistral)	0.1429	0.3619	0.4762	0.1429	0.2251	0.2417	0.1429	0.2592	0.2975	+0.02

Table 3: MiniLM BiEncoder Performance

Experiments - Training with Negative Data

- Adding negative data overall yields better performance
- **Dynamic mining negative data** strategy has significant improvement
  - `Hit@1` +0.039 in BGE model
  - `Hit@1` +0.095 in MiniLM model



Model	Hit@1	Hit@5	Hit@10	MRR@1	MRR@5	MRR@10	NDCG@1	NDCG@5	NDCG@10
BM25	0.1524	0.3143	0.4095	0.1524	0.2184	0.2302	0.1524	0.2425	0.2724
BGE w/o HN	0.2476	**0.4476**	**0.5619**	0.2476	0.3221	0.3373	0.2476	0.3535	0.3904
BGE w/ static HN	0.2286	**0.4476**	0.5333	0.2286	0.3089	0.3219	0.2286	0.3434	0.3726
BGE w/ dynamic HN	**0.2857**	0.4190	**0.5619**	**0.2857**	**0.3352**	**0.3546**	**0.2857**	**0.3562**	**0.4027**
MiniLM w/o HN	0.1714	0.3619	0.4762	0.1714	0.2479	0.2621	0.1714	0.2767	0.3126
MiniLM w/ static HN	0.1714	0.3714	0.4667	0.1714	0.2498	0.2615	0.1714	0.2804	0.3102
MiniLM w/ dynamic HN	**0.2667**	**0.4095**	**0.5333**	**0.2667**	**0.3192**	**0.3345**	**0.2667**	**0.3416**	**0.3804**

Table 5: Negative Training Data Performance on LLaMA Rewrite Keyword

Conclusion

- LLM can perform session and user preference understanding
- Query rerwrite requires suitable prompt template
- Keyword expansion is more effective than directly rewrite
- Utilizing negative data in retrieval model training could significantly boost the overall performance
- Dynamic mining negatives with BiEncoder model achieves best result