

# Data Fusion of Synthetic Query Variants With Generative Large Language Models

ACM SIGIR Conference on Research and Development in Information Retrieval  
in the Asia Pacific Region (SIGIR-AP '24)  
December 9–12, 2024, Tokyo, Japan

**Timo Breuer<sup>1</sup>**

<sup>1</sup>TH Köln - University of Applied Sciences, Cologne, Germany

December 12, 2024

Technology  
Arts Sciences  
TH Köln

# Motivation

## **Alaofi et al. [1]:**

💬 Instruction-tuned large language models can generate user query variants.

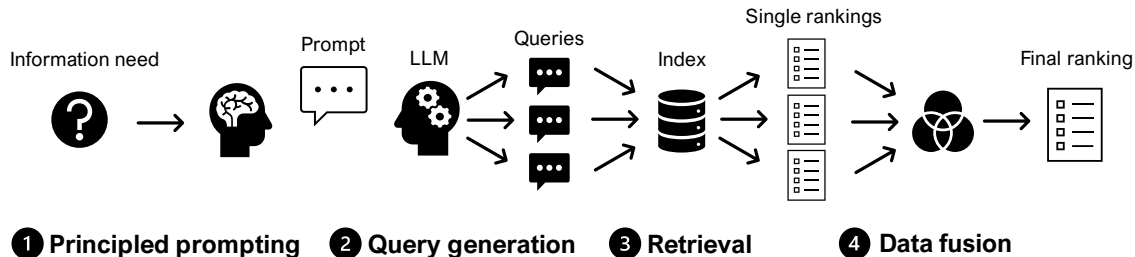
## **Benham and Culpepper [2], Belkin et al. [3]:**

✂ Data fusion with user query variants improves retrieval effectiveness.

## **This work:**

💬 + ✂ Can large language models synthesize effective query variants for data fusion?

# Methodology



# TREC topic 336

<top>

<num> Number: 336

<title> Black Bear Attacks

<desc>

A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.

<narr>

It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

</top>

# Prompting the large language model

## Prompt template for strategies P-1, P-2, and P-3

You are a generator of search query variants.

Generate one hundred keyword queries about <title>.

<description> <narrative>

Example queries for the topic about <example title> include <1st query example>, <2nd query example>, <3rd query example> ...

Your reply is a numbered list of search queries.

P-1:  P-2:  P-3: 

## Prompting strategy P-1:

### Prompt example for topic 336 with strategy P-1

You are a generator of search query variants.

Generate one hundred keyword queries about black bear attacks.

Your reply is a numbered list of search queries.

## Prompting strategy P-2:

### Prompt example for topic 336 with strategy P-2

You are a generator of search query variants.

Generate one hundred keyword queries about black bear attacks.

A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior. It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

Your reply is a numbered list of search queries.

## Prompting strategy P-3:

### Prompt example for topic 336 with strategy P-3

You are a generator of search query variants.

Generate one hundred keyword queries about black bear attacks.

Example queries for the topic about "recycling lead acid batteries" include "1. battery recycling facilities", "2. car battery", "3. car battery disposal". Other query examples for the topic about "symptoms of heart attack" include "1. Early signs and symptoms of having a heart attack", "2. general heart attack symptoms", "3. Heart and stroke foundation", query examples for the topic about "evidence for evolution" include "1. acceptance of theory of evolution", "2. current evidence about the theory of evolution", and "3. current evidence for evolution", ...

Your reply is a numbered list of search queries.



# Query generation with OpenAI's GPT

💬 OpenAI's GPT-4o

⚙️ Fixed random seed, temp. of 0.0

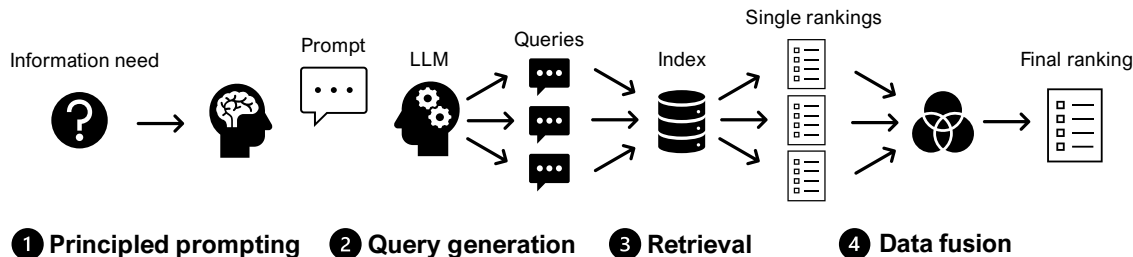
🗄️ Dataset of 40,000 queries

💰 Approx. costs of US\$22.50

## Examples generated by P-1

1. Causes of black bear attacks
2. Black bear attack statistics
3. How to survive a black bear attack
4. Black bear attack prevention tips
5. Black bear attack stories
6. Black bear attack videos
7. Black bear attack news
8. Black bear attack fatalities
9. Black bear attack injuries
10. Black bear attack frequency
- ...

# Retrieval method and data fusion



P-1: 

P-2: 

P-3: 

GPT-4o

**BM25** [4]

**RRF** [5]

# Test collections

**Robust04** TREC Disks 4 & 5 (minus CR)<sup>1</sup> used as part of TREC Robust 2004

**Robust05** AQUAINT Corpus of English News Text<sup>2</sup> used as part of TREC Robust 2005

**Core17** New York Times Annotated Corpus<sup>3</sup> used as part of TREC Common Core 2017

**Core18** TREC Washington Post Corpus<sup>4</sup> used as part of TREC Common Core 2018

---

<sup>1</sup><https://trec.nist.gov/data/cd45/>







<sup>2</sup><https://catalog.ldc.upenn.edu/LDC2002T31>

<sup>3</sup><https://catalog.ldc.upenn.edu/LDC2008T19>







<sup>4</sup><https://trec.nist.gov/data/wapost/>

# Retrieval effectiveness

Comparisons based on **Core17** and **Core18**.

#	Prompt	P@10	NDCG@10	BPref	MAP
Core17					
a	BM25	0.458	0.372	0.274	0.199
b	BM25 + RM3	0.534 <sup>a</sup>	0.404	0.317 <sup>a</sup>	0.246 <sup>a</sup>
c	P-1 	0.526 <sup>a</sup>	0.426 <sup>a</sup>	0.355 <sup>ab</sup>	0.240 <sup>a</sup>
d	P-2 	<b>0.618<sup>ac</sup></b>	<b>0.522<sup>abc</sup></b>	<b>0.416<sup>abce</sup></b>	<b>0.299<sup>abce</sup></b>
e	P-3 	0.570 <sup>a</sup>	0.453 <sup>a</sup>	0.376 <sup>ab</sup>	0.255 <sup>a</sup>
Core18					
a	BM25	0.426	0.389	0.253	0.191
b	BM25 + RM3	0.448	0.396	0.282 <sup>a</sup>	0.229 <sup>a</sup>
c	P-1 	0.452	0.433	0.294	0.233 <sup>a</sup>
d	P-2 	<b>0.532<sup>ace</sup></b>	<b>0.497<sup>abce</sup></b>	<b>0.339<sup>abc</sup></b>	<b>0.270<sup>ac</sup></b>
e	P-3 	0.440	0.420	0.312 <sup>a</sup>	0.240 <sup>a</sup>

Comparisons based on **Robust04** and **Robust05**.

#	Prompt	P@10	NDCG@10	BPref	MAP
Robust04					
a	BM25	0.410	0.421	0.241	0.228
b	BM25 + RM3	0.443 <sup>a</sup>	0.443 <sup>a</sup>	0.268 <sup>a</sup>	0.262 <sup>a</sup>
c	P-1 	0.455 <sup>a</sup>	0.471 <sup>a</sup>	0.269 <sup>a</sup>	0.256 <sup>a</sup>
d	P-2 	<b>0.514<sup>abce</sup></b>	<b>0.535<sup>abce</sup></b>	<b>0.304<sup>abce</sup></b>	<b>0.295<sup>abce</sup></b>
e	P-3 	0.459 <sup>a</sup>	0.480 <sup>ab</sup>	0.276 <sup>a</sup>	0.265 <sup>a</sup>
Robust05					
a	BM25	0.352	0.296	0.227	0.176
b	BM25 + RM3	0.408	0.329	0.251	0.209 <sup>a</sup>
c	P-1 	0.396	0.343	0.267 <sup>a</sup>	0.206
d	P-2 	<b>0.522<sup>ac</sup></b>	<b>0.442<sup>a</sup></b>	<b>0.361<sup>abce</sup></b>	<b>0.267<sup>ac</sup></b>
e	P-3 	0.492 <sup>ac</sup>	0.404 <sup>ac</sup>	0.303 <sup>abc</sup>	0.236 <sup>ac</sup>

Superscripts denote significant differences in paired Student's t-test (with Bonferroni correction applied).

# Information in the topic files is key for effective query generation

## P-1

1. Causes of black bear attacks
2. Black bear attack statistics
3. How to survive a black bear attack
4. Black bear attack prevention tips
5. Black bear attack stories
6. Black bear attack videos
7. Black bear attack news
8. Black bear attack fatalities
9. Black bear attack injuries
10. Black bear attack frequency

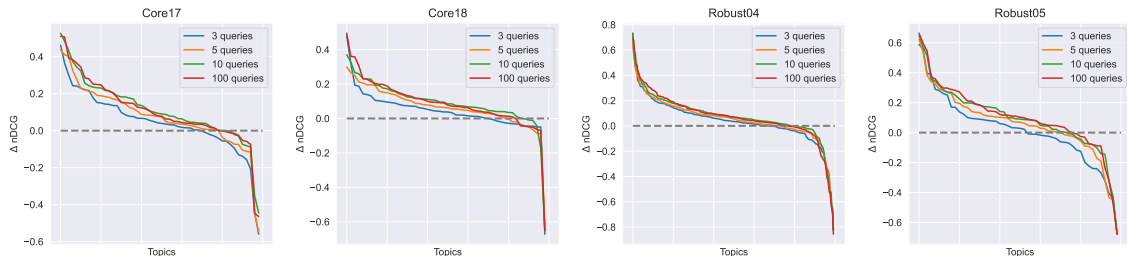
## P-2


1. Frequency of black bear attacks worldwide
2. Causes of black bear attacks on humans
3. Black bear attacks due to food attraction
4. Cosmetics attracting black bears
5. Scientific speculation on black bear attacks
6. Wildlife officials' methods to control black bear attacks
7. New methods to modify black bear behavior
8. Vicious black bear attacks statistics
9. Black bear attacks and human fatalities
10. Preventing black bear attacks

## P-3

1. Black bear attack statistics
2. Black bear attack prevention
3. Black bear attack survival stories
4. Black bear attack frequency
5. Black bear attack locations
6. Black bear attack injuries
7. Black bear attack fatalities
8. Black bear attack news
9. Black bear attack videos
10. Black bear attack photos

# Different numbers of query variants



Retrieval effectiveness with different numbers of synthetic queries based on **P-2** . The plots show the relative improvements in terms of  $\Delta$  nDCG. For each topic of the fused rankings, the difference to the baseline (BM25 with the topic's title as the query) is determined with the four newswire benchmarks Core17/18 and Robust04/05.

# Discussion and wrap-up

*In a nutshell*

➤ **Information in the topic files** is key for effective query generation

# Discussion and wrap-up

*In a nutshell*

- **Information in the topic files** is key for effective query generation

*Putting it into practice!*

- **Conversational agent** obtains context information from user
- **Relieve users** from formulating effective queries and **support RAG systems**



# Discussion and wrap-up

*In a nutshell*

- **Information in the topic files** is key for effective query generation

*Putting it into practice!*

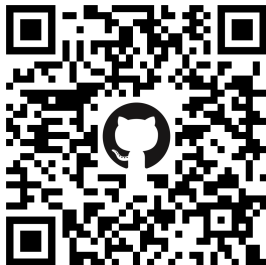
- **Conversational agent** obtains context information from user
- **Relieve users** from formulating effective queries and **support RAG systems**

*Future work*

- Better understanding of **query bias** and **effectiveness tradeoffs**
- Beyond data fusion: **user simulations**, pooling by **query polyrepresentation**

# Thank you for your attention!

Code and data



<https://github.com/breuert/sigirap24>

Pre-print of the paper



<https://arxiv.org/abs/2411.03881>

# References I

- [1] M. Alaofi, L. Gallagher, M. Sanderson, F. Scholer, and P. Thomas, “Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study,” in *SIGIR*, ACM, 2023, pp. 1869–1873.
- [2] R. Benham and J. S. Culpepper, “Risk-Reward Trade-Offs in Rank Fusion,” in *ADCS*, ACM, 2017, 1:1–1:8.
- [3] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan, “Effect of Multiple Query Representations on Information Retrieval System Performance,” in *SIGIR*, ACM, 1993, pp. 339–346.
- [4] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, “Okapi at TREC-3,” in *TREC*, ser. NIST Special Publication, vol. 500-225, National Institute of Standards and Technology (NIST), 1994, pp. 109–126.
- [5] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,” in *SIGIR*, ACM, 2009, pp. 758–759.