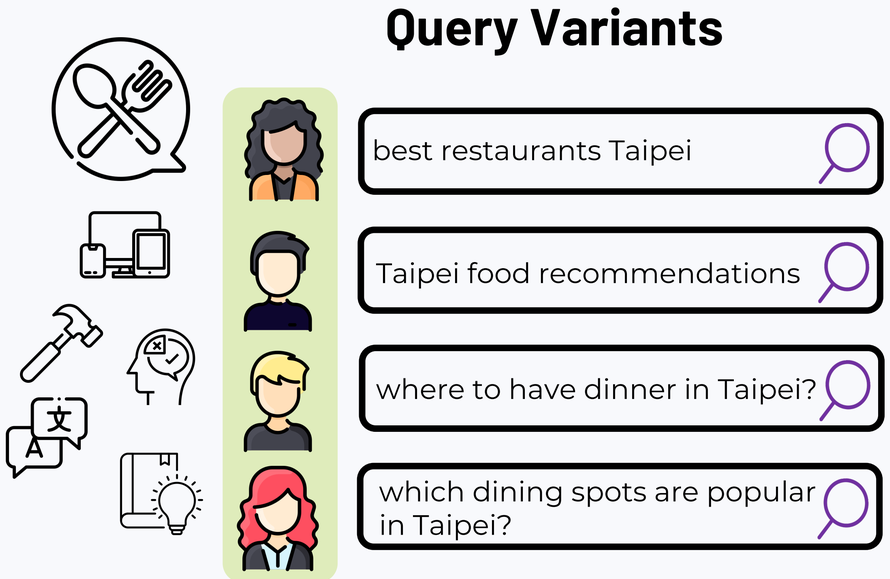
# Can Generative LLMs Create Query Variants for Test Collections?

## *An Exploratory Study*

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, Paul Thomas

## Query Variants

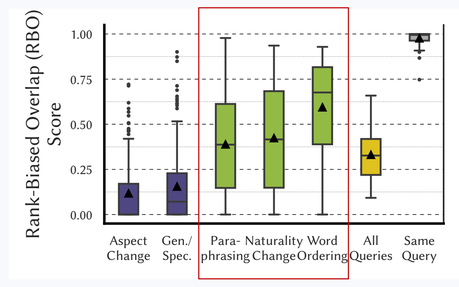



## Why Do Variants Matter?

**(1) Retrieval Performance[1]**
*Consistency and effectiveness*

**(2) Document retrievability (document exposure)[1]**

**(3) Pooling in test collections**




Inconsistent result pages

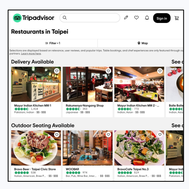
[1] *Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From? In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2850–2862. https://doi.org/10.1145/3477495.3531711*

## 1. Query Variant Generation

### (1) Using Query Logs

Assuming a shared click indicates a shared intent

best restaurants Taipei
where to eat in Taipei
⋮
food in Taipei




### (2) Crowd-sourcing

Prompting users to generate query variants

backstory/information need description

### (3) Query simulation

## 2. Research Questions

**RQ1:** Can an LLM, with one-shot learning, generate queries that are similar or perhaps identical to human ones?

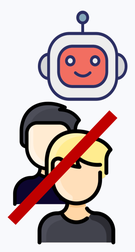
**RQ2:** How do the human queries and the LLM queries compare when used for document pooling when constructing test collections?

## 3. Method

**Prompting ~~Users~~ LLMs**

Example

*You have an important presentation to make, and decide to wear a jacket and tie. You know that the "windsor knot" is recognized as being the most stylish way of tying a tie, but have no idea how to do one, and would like to find out.*



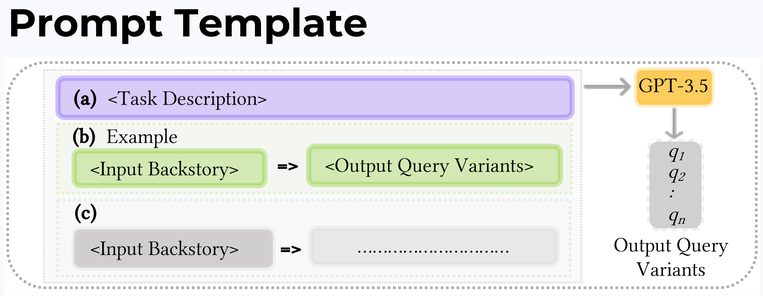
**Human-generated queries**

1. **how to tie a windsor knot**
2. instructions for tying a windsor knot
3. windsor knot directions
4. youtube windsor knot how to tie a tie
5. how do i make a windsor knot

**GPT-3.5-generated queries**

1. **how to tie a windsor knot**
2. windsor knot tutorial
3. windsor knot how to
4. windsor tie knot tying instructions
5. what is a windsor knot

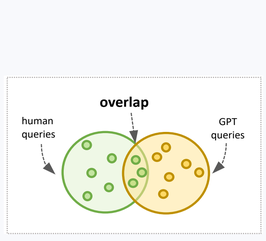
**Prompt Template**

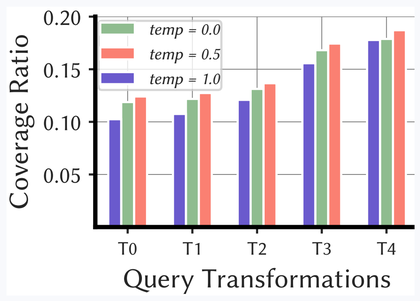



## 4. Evaluation Metrics and Results

### Are LLMs capable of generating query variants?

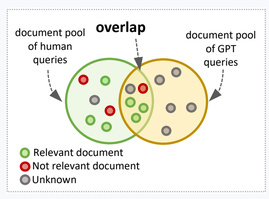
**RQ1:** Similarity to Human-generated Queries

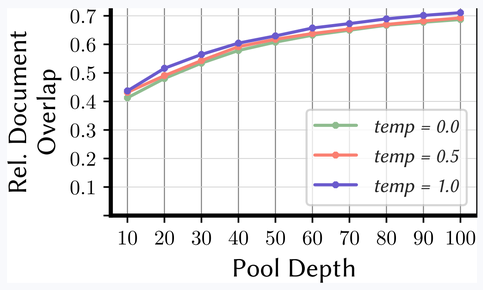



~18% re-produced queries

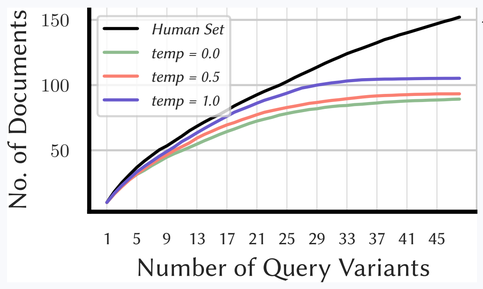
**RQ2:** Document Pool Similarity and Effectiveness Scores




~70% overlapped documents

The average *pool size* at depth 10 as more variants are added.




Higher growth in the human query pool

Average *effectiveness* metrics, *RBO* and *pool properties* at depth 10

| Variant Set | P@10 | NDCG@10 | RBP | RBO | Pool Properties | | |
|---|---|---|---|---|---|---|---|
| | | | | | Size | Relevant | Unjudged |
| Human set | 0.443 | 0.274 | 0.406 +0.111 | 0.201 | 190.69 | 0.30 | 0.13 |
| GPT (*temp* = 0.0) | 0.386‡ | 0.246† | 0.358 +0.254‡ | 0.235† | 94.42 | 0.29 | 0.33 |
| GPT (*temp* = 0.5) | 0.393‡ | 0.249† | 0.360 +0.238‡ | 0.220 | 93.55 | 0.29 | 0.31 |
| GPT (*temp* = 1.0) | 0.384‡ | 0.240‡ | 0.355 +0.263‡ | 0.235† | 105.21 | 0.27 | 0.37 |

## 5. Conclusions and Future Directions

- ❖ An opportunity to expand queries in existing test collections.
- ❖ Further research to compare the use of LLMs with previous methods of query simulation.

Marwah Alaofi [13], Luke Gallagher[1], Mark Sanderson[13], Falk Scholer[13], Paul Thomas[2]
[1]RMIT University, [2]Microsoft, [3]ARC Centre of Excellence for Automated Decision Making + Society

*Marwah Alaofi is supported by a scholarship from Taibah University, Saudi Arabia. This work was also supported by the Australian Research Council (DP180102687, CE200100005).*