Semantic Search Evaluation

Chujie Zheng[*]
LinkedIn
Sunnyvale, California, USA

Jeffrey Wang
LinkedIn
Sunnyvale, California, USA

Shuqian Albee Zhang
LinkedIn
Sunnyvale, California, USA

Anand Kishore
LinkedIn
Sunnyvale, California, USA

Siddharth Singh[†]
Walmart
Sunnyvale, California, USA

ABSTRACT

We propose a novel method for evaluating the performance of a content search system that measures the semantic match between a query and the results returned by the search system. We introduce a metric called "on-topic rate" to measure the percentage of results that are relevant to the query. To achieve this, we design a pipeline that defines a golden query set, retrieves the top K results for each query, and sends calls to GPT 3.5 with formulated prompts. Our semantic evaluation pipeline helps identify common failure patterns and goals against the metric for relevance improvements.

KEYWORDS

Semantic Relevance, Search System Evaluation, Content Search, Information Retrieval, Generative AI

1 INTRODUCTION

LinkedIn search has made significant progress over the years by incorporating semantic matching capabilities. Semantic matching capability helps members find knowledge more easily by delivering results that are conceptually related to their search query, even if they do not contain the exact keywords used in the query. As we continue to improve our system, we strive to overcome the complex quality challenges that arise:

- Indirect measurement: While we have existing engagement metrics from online experiments, they do not necessarily capture the quality of our search results. When member feedback comes in, there is an added overhead in determining if the system is operating as expected.
- Not operationalized: Search patterns and expectations change over time, so we need an automated way to continuously measure quality.

To address these gaps, we present a semantic evaluation pipeline that leverages Generative AI (GAI) to evaluate quality. Our main contributions and business impacts are:

- We propose the metric "on-topic rate" to measure the percentage of content search results that are relevant to the intended topic. This metric serves as a tool for evaluating the performance of the content search model, and can also help identify common failure patterns. By setting goals based on this metric, we can work towards improving the relevance of search results.
- We present a novel semantic evaluation pipeline for search engine offline evaluation. This framework helps translate user problems into technical patterns that can be operationalized and plays a critical role for search engine offline evaluation.
- We present two approaches to evaluate the Generative AI evaluation output, including conducting human evaluation and preparing a validation set to ensure the quality.

2 RELATED WORK

Assessing the quality of search engine results has become a challenging task, particularly in determining which ranking models perform best under specific business metrics [2, 8]. The most reliable method for evaluating model performance is through online A/B testing, but this approach has limitations [5, 7]. Firstly, due to the high cost of traffic, only a limited number of models can be compared within a given timeframe, as it requires a significant amount of user feedback to draw statistically significant conclusions. Secondly, there is a risk of negatively impacting the user search experience if the model performs poorly. Consequently, offline evaluation is commonly used to select candidates for online experiments.

Existing work has proposed a variety of techniques for search engine offline evaluation, including relevance, reliability, timeliness, diversity and fairness [3]. Commonly we evaluate the quality of search by using the evaluation metrics focused on relevance, such as Mean Average Precision (MAP) and normalized Discounted cumulative gain (nDCG). These types of metrics simulate how users interact with the search engine result page (SERP) [4, 14]. Nonetheless, a prevalent drawback of offline evaluation lies in its dependence on historical data, and these metrics may not directly measure the quality of search engine, which can result in imperfect correlations with users' search experiences [1, 6, 10, 16].

Recent advancements in semantic search and generative AI have significantly transformed the landscape of search systems, particularly with the integration of large language models (LLMs) in ranking and recommendation tasks [9, 15, 17]. LLMs are also explored for relevance judgments and evaluation in information retrieval [12]. The literature has shown that LLMs can closely replicate human judgements through prompt strategies [13] and highlights the growing role of LLMs in automated evaluations [11].

3 ON-TOPIC RATE

In this section, we formulate our task and introduce our proposed metric - On-Topic Rate (OTR). This metric is proposed as the direct measurement to evaluate if the retrieved document is primarily about the query.

[*] Corresponding author, email: razheng@linkedin.com.
[†] Work done in LinkedIn.

3.1 Task Formulation

Given a member query q as input, the search engine returns a list of documents $D = (d_1, d_2, ..., d_n)$. Since we focus on content search, here each document corresponds to a post or article on LinkedIn.

3.2 Computation

On-topic rate is a metric that measures the relevance of search results to the user's query. This metric helps evaluate how well the search engine is able to understand the query's intent and provide relevant results. A high on-topic rate indicates that content search is performing well and providing useful results to our members, while a low on-topic rate suggests that improvements may be needed to better understand the user's search intent. We define OTR for <query, doc> pair as the following:

$$OnTopicRate(q, d_i) = 1 \text{ if the pair is relevant, otherwise } 0 \quad (1)$$

3.2.1 *OTR@K.*

We measure the on-topic rate for the ML model by selecting the top K returned documents for each query. We define OTR@K as the total number of query-document pairs that are relevant divided by the total number of returned documents.

$$OTR@K = \frac{\sum_{i=1}^{K} OTR(q, d_i)}{K} \quad (2)$$

4 SEMANTIC EVALUATION SETUP

Within this section, we detail the process of establishing semantic evaluation and harnessing Generative AI (GAI) to discern the relevance between queries and documents. Figure 1 has illustrated the overall pipeline: our methodology begins with the creation of a test query set, followed by the formulation of prompts provided to the Large Language Model (LLM) and subsequent calculation of OTR based on the GAI-generated outputs.

4.1 Create Query Set

We construct the query set used for evaluation by leveraging different resources to comprehensively cover the relevant topics or areas of interest.

4.1.1 *Golden Set.*

The golden set serves as a stable and uniform standard for assessing and benchmarking queries. It includes the aspirational queries and the top queries from production. We include the following types of query into the golden set:

- Top queries: We incorporate common queries, which can be seen as the most popular and indicative keywords that members are searching for.
- Topical queries: Topical queries encompass search inquiries or questions that pertain to specific subjects or topics. We incorporate these into our golden set because they pose a greater evaluation challenge, often being lengthier and more intricate in their intent. Providing high-quality results for topical queries can enhance our members' ability to access valuable knowledge on LinkedIn.

4.1.2 *Open Set.*

The Open Set is a dynamically changing set of queries used for evaluation. The source of the open set includes:

- trending queries and newsy queries from production
- some random queries from production, to add diversity

	Example Queries
Top queries	covid-19, resume, microsoft excel, we're hiring, work from home
Topical queries	how to create a personal brand, how to stand out in a competitive job market, how do I negatiate my salary

Table 1: Example golden set queries from different sources

Table 2 are some example queries from the golden set and open set used in production for evaluation.

	Example Queries
Golden Set	improve workplace communication, remote team best practices, how do I get promoted
Open Set	fed raises rates,leadership first, barbie, women ai study

Table 2: Example queries in golden set and open set

4.2 Get search results for query set

We gather the following information for each query within the specified set: the top $K = 10$ documents from production, along with post textual information and any mentioned article information (including title and article body, if applicable).

4.3 Formulate the prompt

We construct the prompt for GPT 3.5 to collect feedback from the LLM. Using the inputs collected from previous steps, we define the prompt to encompass three perspectives:

1. the definition of On-Topic rate
2. detailed guidance for decision making
3. query
4. post, including all the text information from the posts, including the commentary and any re-shared posts/articles

4.3.1 *Metric Definition.*

We begin by defining a metric and then task LLM with making decisions based on our specific requirements. We keep iterating our prompts to improve the accuracy of the decision. Based on our experiences from production and our prompting practices, we have learnt that providing very precise definition for the request and adding examples with provided reasons can improve the performance significantly.

For example, for our task, we tested two prompts in production for our use case:

- **Prompt A**: Given the post below, is the post strongly relevant to the query?
- **Prompt B**: Given the post below, is the post primarily about query or strongly relevant to the query?

Despite the subtle differences between the two prompts, Prompt B outperforms Prompt A significantly. This is because Prompt B specifically directs attention to the main subject of both the query and the post. As a result, it reduces the false positive that might

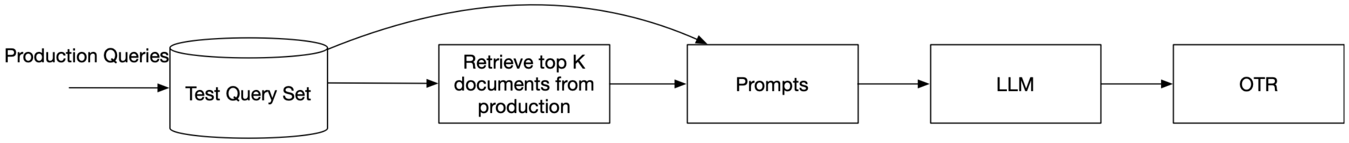

Figure 1: Semantic evaluation pipeline

occur due to keyword matching. We have included some examples to show the impacts in Appendix A.

We discover that incorporating well-detailed examples alongside the reasons behind the decision can enhance performance and address various problems identified from observed patterns of failure. With no examples in the prompt the results did not align with human judgment. By summarizing the failure pattern, during our prompt iteration, we determine that incorporating examples with explanations of the judgments can boost the generative AI's capacity for better decision making.

4.3.2 Guidance. The guidance offers comprehensive instructions for training LLM to make decisions that align with our expectations. This may involve outlining decision criteria and specific requirements, as well as providing examples of corner cases to illustrate decision-making processes.

Here is an example of defining guidance for OTR:

(1) The on-topic decision should not only consider the keyword match between query and post. It should reflect the semantic matching between query intention and the post details.
(2) The post information is primarily relevant to the user query.

4.4 Compute OTR Metrics

From the GAI output, we have three types of information:

- Binary decision: It directly corresponds to if the retrieved post is primarily on-topic of the query.
- Relevance score: It is the score related to the binary decision. The score is in the range of 0 and 1. The relevance score aims to measure the semantic relevance between query and post. The decision should keep the consistency between the relevance score.
- Decision reason: It explains the reason for the binary decision and relevance score. This field is not used for OTR calculation but it is very important to help us iterate the prompt, since it explains the consideration for the GAI decision.

Table 3 presents an instance from the pipeline where a post mentioning an article introducing tips for self-promotion was retrieved for the query "promotion tips". Although this example may be classified as on-topic through keyword matching, it does not accurately reflect the query intention. To address this issue, semantic evaluation enables the identification of the primary topic of the post. Based on the decision reason, it can be concluded that the prompt provided instructions for measuring semantic relevance, and the LLM was able to identify that although the keyword matched the query, the post itself did not address the query intention, resulting in a low relevance score and an off-topic binary decision.

Our approach consider (query, document) as on-topic only if the binary decision is 1 and the relevance score exceeds a certain threshold (we use 0.5 in production after analyzing the distribution of relevance score). This ensures that we only consider pairs that the GAI has a strong confidence in, and disregard those that fall below the threshold. By taking this rule, we compute *OTR@K* and nDCG as the final output.

Query	promotion tips
Post	Here are 13 tips to get you over that mental hurdle. #speakup #tips #leadership
Binary decision	0
Relevance score	0.4
Decision Reason	The post is about tips for self-promotion and personal branding, and does not directly address the query of "promotion tips". The mention of "tips" in the hashtags is not specific enough to make a strong connection to the query.

Table 3: Example output from semantic evaluation pipeline.

5 EXPERIMENT

5.1 Human Evaluation on Generated Output

We employ human evaluation as a metric to ascertain the alignment between decisions made by the proposed semantic evaluation pipeline and expert human annotators, aiming to verify the reliability and consistency of the proposed pipeline in delivering high-quality results.

Our human evaluation team comprises 10 colleagues, and all team members possess substantial experience in content search to ensure the credibility of their decisions. To maintain the quality of the annotation process, annotators underwent a rigorous qualification process. This process included familiarizing themselves with annotation guidelines and assessing 20 representative query-response pairs. Following this, we conducted individual assessments of the annotations submitted and organized group discussions to address any confusions or uncertainties regarding the task. Each annotator was given the task of annotating 50 pairs of queries and documents. Both annotators and the GAI are providing the same information: query, post commentary, and re-shared posts/articles.

To gauge the level of agreement among annotators, we collected annotations for a specific set of randomly selected query-response pairs, which yielded a high degree of agreement among the annotators.

Evaluating a query-documents pair requires our annotators to complete a three-step evaluation:

(1) Evaluate if the query and the post is relevant

(2) Provide reasons regarding the judgement if it is considered as irrelevant

We utilize the (query, document) data from the annotation result as input for the proposed semantic evaluation process and then compare the output of the pipeline with human evaluation to assess its consistency. By comparing the results, we report 81.72% consistency with the GAI decision.

5.2 Performance on Validation Set

To ensure the quality of our prompt, the team has compiled a validation set comprising various query types and corresponding post pairs, serving as the baseline for assessing the prompt's effectiveness. We select representative queries from production that are frequently searched by members, including:

- company name queries
- title queries, like data engineer, product manager
- skill queries, like finance, customer services, marketing
- newsy queries, like february jobs report
- other top queries, for example: work from home, open to work

Total our validation set includes 60 queries, each paired with 10 related posts, resulting in 600 query-post pairs. Our team assesses these pairs and makes a binary determination regarding their topical relevance. This judgment is confirmed by at least three team members. We use this validation set as our standard of truth for gauging the effectiveness of our prompts. Furthermore, we utilize this validation set to identify common failure patterns to keep iterating our prompts. The current prompt used in production achieves 94.5% accuracy on the validation set.

6 HOW DO WE USE SEMANTIC EVALUATION TO IMPROVE THE PRODUCT?

We have adopted the semantic evaluation pipeline as our offline benchmark for experiments in LinkedIn content search. This serves as the foundation for offline evaluation, measuring whether the trained ML model has correctly captured the query intention. The pipeline is designed to monitor the performance of the served content search model, with a weekly update of the open query set and pipeline run to ensure that the calculated OTR falls within the desired range. We also use this tool to evaluate new trained ML models offline, comparing them to the baseline to quickly test for improvements and provided feedback for iteration.

In addition, we explore the decision reasons to identify growth opportunities. We collect cases identified as off-topic from the pipeline and identify performance gaps that we can improve.

7 CONCLUSION

In this work, we introduce a semantic evaluation pipeline for search engine offline evaluation. Our proposed metric, on-topic rate, measures the relevance of search results to the user's query. We also outline the construction of prompts and calculation of the OTR, and demonstrate the high consistency between human evaluation and pipeline output. With this semantic evaluation framework, we are able to directly measure the quality of post-search results and gain a better understanding of our search engine's performance.

ACKNOWLEDGMENTS

The authors would like to thank Rupesh Gupta, Raghavan Muthuregunathan, Dhruv Saksena, Abhi Lad, Alice Xiong, Rashmi Jain for their support of the work. We also acknowledge and thank the contributions and feedback from our colleagues in LinkedIn Content Search AI team: Xin Yang, Madhumitha Mohan, Ali Hooshmand, Justin Zhang and Sarang Metkar.

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* 773–774.

[2] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-objective ranking optimization for product search using stochastic label aggregation. In *Proceedings of The Web Conference 2020.* 373–383.

[3] Nuo Chen, Donghyun Park, Hyungae Park, Kijun Choi, Tetsuya Sakai, and Jinyoung Kim. 2023. Practice and Challenges in Building a Business-oriented Search Engine Quality Metric. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 3295–3299.

[4] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval.* 15–24.

[5] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[6] Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* 567–574.

[7] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1517–1525.

[8] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On application of learning to rank for e-commerce search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval.* 475–484.

[9] Xiaoxi Li, Zhicheng Dou, Yujia Zhou, and Fangchao Liu. 2024. Corpuslm: Towards a unified language model on corpus for knowledge-intensive tasks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 26–37.

[10] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does relevance mean usefulness and user satisfaction in web search?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* 463–472.

[11] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Llm4eval: Large language model for evaluation in ir. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 3040–3043.

[12] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Report on the 1st Workshop on Large Language Model for Evaluation in Information Retrieval (LLM4Eval 2024) at SIGIR 2024. *arXiv preprint arXiv:2408.05388* (2024).

[13] Hossein A Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles LA Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. LLMJudge: LLMs for Relevance Judgments. *arXiv preprint arXiv:2408.08896* (2024).

[14] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[15] Zhizhong Wan, Bin Yin, Junjie Xie, Fei Jiang, Xiang Li, and Wei Lin. 2024. LARR: Large Language Model Aided Real-time Scene Recommendation with Semantic Understanding. *arXiv preprint arXiv:2408.11523* (2024).

[16] Xiaojie Wang, Ruoyuan Gao, Anoop Jain, Graham Edge, and Sachin Ahuja. 2023. How well do offline metrics predict online performance of product ranking models?. In *SIGIR 2023.* https://www.amazon.science/publications/how-well-do-offline-metrics-predict-online-performance-of-product-ranking-models

[17] Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference*

on Research and Development in Information Retrieval. 38–47.

A EXAMPLES TO DEMONSTRATE THE PERFORMANCE IMPROVEMENT BETWEEN PROMPT A AND PROMPT B

During our experiments, we have noticed the importance of formatting the request into very precise guidance that can help GAI understand the request and leverage it for decision making. For example, for two prompts mentioned earlier, given the same information for query and posts, we are seeing Prompt B outperforms Prompt A by significantly reducing the false positives caused by keyword matching.

- **Prompt A**: Given the post below, is the post strongly relevant to the query?
- **Prompt B**: Given the post below, is the post primarily about query or strongly relevant to the query?

Here are some examples that Prompt A and Prompt B are making different decisions in Table 4.

Query	Post	Prompt A	Prompt B
react native	Hey LinkedIn community, In my recent article, "The Rising of New Gen-GPT and Socializing as an Adult," I dove into some exciting topics that I want to share with you in a nutshell: Projects in Progress: I've been working on various projects, including exploring the Spotify API for a dynamic playlist-based website and a new React Native project. The latter aims to address real-world needs, from allergen tracking to task delegation tools and a notes app integrated with Slack to help in my current position. AI Evolution: The latest Gen-GPT update is a game-changer Balancing Act: I shared my personal reflections on the importance of work-life balance. While pursuing my career, I realized the value of reconnecting with friends, especially those from my high school days, as a source of relaxation and genuine joy. These are just the highlights. If you want to dive deeper into these topics, check out the full article linked in this post. Your feedback and insights are always appreciated. Let's keep the conversation going! I write them weekly it is something will not want to miss.	On Topic	Off Topic
manager	Twin Transformation is currently THE topic of sustainability managers, but also of digitalisers. Brigitte Falk was a pioneer in this field 20 years ago. So her interview (Digitization as a bridge to sustainable business) is hopefully an inspiration and encouragement for many as a look back, but above all as a look forward!	On Topic	Off Topic
best practices for managing remote teams	Excited to invite you to my upcoming workshop at #ReactIndia 2023 with my co-speaker Lokesh Yadav on "Building a Web Performance Culture: Empowering Large-Scale Teams to Deliver Lightning-Fast User Experiences"! In today's digital world, web performance plays a crucial role in user experience, business success, and stakeholder satisfaction. Slow-loading websites can significantly impact user engagement, conversion rates, and revenue. That's why it's essential for large-scale teams to establish a web performance culture and empower their teams to deliver exceptional user experiences. Join me for this workshop where we'll dive deep into the key aspects of building a web performance culture within larger teams. We'll explore the significance of performance for you and your stakeholders, and learn how to differentiate between noise and reality by implementing the right tooling for Single-Page Applications (SPA) and Multi-Page Applications (MPA). Attendees will gain valuable insights into improving performance metrics and discover innovative approaches using Real User Monitoring (RUM) and Lab data to track and monitor performance enhancements without incurring additional costs. We'll also cover practical strategies, best practices, and real-life examples, including tackling performance challenges with React 18 Hydration. Whether you're a developer, manager, or part of a large-scale team, this workshop will help you to build new perspectives towards web performance and empower you to deliver lightning-fast user experiences.	Off Topic	On Topic

Table 4: Examples from Prompt A and Prompt B that are making different decisions.