Towards More Relevant Product Search Ranking Via Large Language Models: An Empirical Study

Qi Liu
qi.liu@walmart.com
Walmart Global Tech
Hoboken, NJ, USA

Atul Singh
atul.singh@walmart.com
Walmart Global Tech
Hoboken, NJ, USA

Jingbo Liu
jingbo.liu@walmart.com
Walmart Global Tech
Hoboken, NJ, USA

Cun Mu
cun.mu@walmart.com
Walmart Global Tech
Hoboken, NJ, USA

Zheng Yan
zheng.yan0@walmart.com
Walmart Global Tech
Hoboken, NJ, USA

Abstract

Training Learning-to-Rank models for e-commerce product search ranking can be challenging due to the lack of a gold standard of ranking relevance. In this paper, we decompose ranking relevance into content-based and engagement-based aspects, and we propose to leverage Large Language Models (LLMs) for both label and feature generation in model training, primarily aiming to improve the model's predictive capability for content-based relevance. Additionally, we introduce different sigmoid transformations on the LLM outputs to polarize relevance scores in labeling, enhancing the model's ability to balance content-based and engagement-based relevances and thus prioritize highly relevant items overall. Comprehensive online tests and offline evaluations are also conducted for the proposed design. Our work sheds light on advanced strategies for integrating LLMs into e-commerce product search ranking model training, offering a pathway to more effective and balanced models with improved ranking relevance.

CCS Concepts

• **Computing methodologies** → **Learning to rank**; **Ranking**; **Natural language processing**.

Keywords

Product search ranking, Large language models, Learning to rank, Search relevance

ACM Reference Format:
Qi Liu, Atul Singh, Jingbo Liu, Cun Mu, and Zheng Yan. 2024. Towards More Relevant Product Search Ranking Via Large Language Models: An Empirical Study . *Proceedings of the first workshop on Generative AI for E-Commerce 2024, October 25, 2024.* Boise, Idaho, USA, 6 pages.

Genaiecom '24, October 25, 2024, Boise, ID

1 Introduction

As an essential daily activity for millions of users, online shopping has become a crucial element of people's lives. Modern e-commerce product search platforms handle a vast multitude of customer search queries, striving to deliver the most relevant and deserving products in the search results. Product search ranking is one of the most crucial components in the search tech stack, where advanced machine learning models are employed to present online customers with highly relevant products for their search queries. With the rapid advancements in generative artificial intelligence (GenAI) methodologies, particularly large language models (LLMs), there is significant potential to leverage these technologies to improve the relevance of search ranking results. This, in turn, can enhance the system's ability to promote higher-quality products to top positions.

The application of pre-trained language models in search ranking modeling has been widely explored in previous works. For example, as one of the most popular choices, BERT [6] has been extensively used in encoding queries and documents to capture and evaluate semantic relatedness in many studies [12, 24, 25, 40, 42], which achieve state-of-the-art performance on various benchmarks optimizing for ranking relevance. Similar research on search ranking was also performed leveraging variant models of BERT, e.g., DistilBERT [29], ALBERT [15], RoBERTa [21]. With the emergence of more complex LLMs such as LlaMa [36, 37], GPT [2, 26], Mistral [11], and Gemma [33], which are trained on extensive datasets with a larger number of parameters, offering greater flexibility and generalization, and demonstrating improved performance across various semantic tasks [14, 27, 34], it is increasingly tempting to utilize them in search ranking tasks.

Most current methods for integrating LLMs into ranking model training focus on directly using query-product level predicted scores from LLMs as relevance labels for downstream tasks or student models [18, 38, 41]. However, there is less discussion on how to further refine the LLM output for optimal use and how to better integrate LLMs into ranking model training to fully leverage their semantic capabilities. Our work addresses this gap by implementing several key strategies. On the one hand, we integrate LLMs of varying sizes into generating labels and features for ranking model training. While the labels can be generated using more complex

language models, smaller models are used for features due to runtime latency concerns. This approach maximizes improvements in content-based relevance for ranking. On the other hand, we decompose the target for product ranking—relevance labels—into content and engagement components, leveraging LLMs specifically for content relevance. Additionally, we propose several sigmoid-based transformations on the LLM output to generate content relevance, with different transformation choices influencing the dynamics of the engagement vs. content trade-off.

The remainder of the paper is organized as follows: Section 2 discusses our proposed methods for integrating LLMs into training search ranking models. Section 3 details the empirical experiments and evaluations conducted to assess our proposals. Finally, Section 4 summarizes our findings and presents our conclusions.

2 Methodology

Our search ranking models are trained using the Learning-to-Rank (LTR) framework [7, 19, 30] optimized for search events [35] by utilizing data from a truncated historical period of online customer search traffic on Walmart.com, a major e-commerce platform. The improvement of ranking model training is achieved not only through the utilization of high-quality features but also by customizing the labels to guide the model in generating more relevant ranking results that suit specific goals for certain use cases. In this work, our primary focus is on optimizing the integration and utilization of LLMs into both the label and feature generation for model training.

2.1 Label Formulation

We adopt the listwise approach [3] for our LTR modeling, where the loss function aims to optimize the Normalized Discounted Cumulative Gain (NDCG) metrics [10], with relevance scores playing a vital role in the calculation. In the context of e-commerce search, we decompose the ranking relevance into two major components: content-based relevance and engagement-based relevance. The former gauges how pertinent a product's attributes–e.g., title, description, brand, gender, color, product type, etc.–are to a search query, while the latter evaluates how frequently customers interact with a product based on their judgment under a search query. Accordingly, we propose the following label formulation: given a group $\mathcal{G}$ of products $\{p_1, p_2, \ldots, p_{|\mathcal{G}|}\}$, we assign label

$$y^{\mathcal{G}}_{q,p} = \sigma\Big(C_{q,p}\Big) \cdot E^{\mathcal{G}}\Big(q, p\Big), \tag{1}$$

where $C_{q,p} \in [0, 1]$ is an inferred content-based relevance score for query-product pair (q, p), $\sigma(\cdot)$ is a transformation function for the content relevance, and $E^{\mathcal{G}}(\cdot)$ yields the engagement-based relevance score, a value based on the engagement types (ordered > added-to-cart > clicked > non-engaged) for product p in the search event with query q and product group $\mathcal{G}$.

Compared to engagement-based relevance score E, which entirely relies on factual logged customer behavioral data and is therefore objective, content-based relevance C is more subjective, as it depends heavily on semantic interpretation. Given the substantial volume of query-product pairs in e-commerce catalogs, using human-labeled datasets to train machine learning models to predict relevance scores becomes an appealing approach. In this specific instance, we propose to leverage LLM models, distinct from the ranking model, to generate content relevance scores due to their demonstrated exceptional performance on natural language tasks. Specifically, we fine-tune a Mistral 7B model [11] using in-house human-evaluated data with the cross-entropy loss

$$\mathcal{L}_{q,p} = -r_{q,p} \log\Big(\hat{r}_{q,p}\Big) - \Big(1 - r_{q,p}\Big) \log\Big(1 - \hat{r}_{q,p}\Big), \tag{2}$$

where $r_{q,p}$ is the human-evaluated content-based relevance label for query-product pair (q, p) and $\hat{r}_{q,p}$ is the LLM predictions. The fine-tuned LLMs will then be able to infer content-based relevance scores $C_{q,p} \in [0, 1]$ for any given (q, p), where the closer the value is to 1, the more content-relevant the product is.

Though the multiplication of content and engagement can account for both aspects, it also exhibits a potential trade-off between the two, where the engagement performance lift in a ranking model could negatively impact its performance in content-based relevance and vice versa.

2.2 Relevance Transformation

Another perspective for comparing content-based vs. engagement-based relevances is that content reflects endogenous properties of a product, as it originates from inherent attributes of the product itself. In contrast, engagement represents exogenous characteristics, as it depends on how end users perceive and interact with the product. Hence, we let content relevance function as a guardrail, as an ideal ranking model should be able to distinctly differentiate products with significant content relevance gaps, ensuring that highly relevant ones are ranked in top positions. For products within the same content relevance interval, the model should primarily rely on user engagement to determine their ranking. Following this principle, we propose a sigmoid transformation on the content relevance score used in label (1)

$$\sigma(C; \alpha, \beta) = \frac{1}{1 + \exp\big[-\alpha(C - \beta)\big]}, \tag{3}$$

where C is the LLM-inferred content relevance score, and $\alpha, \beta > 0$ are shape parameters determining the center and steepness of the sigmoid curve. The main rationale behind this transformation is that the sigmoid function can polarize those moderately low and high scores toward both extremes, effectively differentiating the content relevance gap between products. Additionally, within each interval at both ends, the content curve remains relatively flat, allowing engagement E to play a more significant role in the label.

Shown in Figure 1 is the comparison between different sigmoid transformation curves. We segment query(q)-product(p) pairs (QPs) into three intervals $\{R_1, R_2, R_3\}$ according to their original content relevance scores $C_{q,p}$ predicted by LLMs. Suppose $0 < c_1 < c_2 < 1$,

- $R_1 = \{(q, p) : 0 \le C_{q,p} < c_1 \text{ s.t. } 0 < \nabla_{C_{q,p}}\sigma \le 1\}$: QPs falling into this interval have low content relevance. They are flattened by the transformation and pushed to have even lower scores. The distinction among these QPs is more determined by their engagement $E^{\mathcal{G}}\big(q, p\big)$.
- $R_2 = \{(q, p) : c_1 \le C_{q,p} < c_2 \text{ s.t. } \nabla_{C_{q,p}}\sigma > 1\}$: QPs falling into this interval have medium content relevance, indicating difficulty in ascertaining their content quality due to

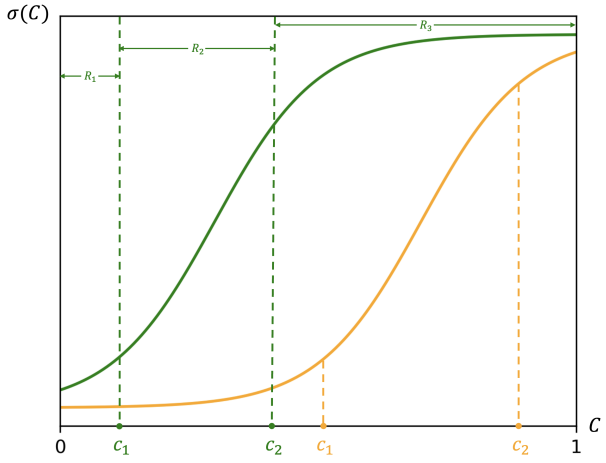

Figure 1: Curve of Sigmoid Transformation

ambiguity. Different customers may exhibit very distinct engagement behaviors, resulting in high variance and reduced predictive power of engagement relevance. The transformation in this area, with the property $\nabla_{C_{q,p}}\sigma > 1$, magnifies the effect of content relevance, compensating for the weakened engagement predictability. This is another benefit of the sigmoid transformation.

- $R_3 = \{(q,p) : c_2 \le C_{q,p} \le 1 \text{ s.t. } 0 < \nabla_{C_{q,p}}\sigma \le 1\}$: QPs falling into this interval have high content relevance. They are flattened by the transformation and pushed to higher scores. The relative orders among these QPs are more determined by their engagement $E^{\mathcal{G}}(q,p)$.

The choices of α and β determine ranges of R_1, R_2, and R_3. The two curves in Figure 1 exhibit the trade-off mainly between R_1 and R_3. The green curve, with a wide range of R_3, sets a low bar for a QP to be highly content-relevant, resulting in more QPs having high content scores. Conversely, the yellow curve, with a narrow range of R_3, sets a high bar for highly content-relevant QPs, leading to only a few of them being assigned high content scores.

In product search ranking, the top few positions, such as the top 5 and top 10, hold the most importance among the entire recall set, which usually contains hundreds or even thousands of products. This contrast emphasizes the need to prioritize those few items with the best content relevance in the top positions. Therefore, having a narrow but high-quality R_3 interval could potentially benefit the content relevance. This approach ensures the content quality of the R_3 products that are likely to occupy most of the top positions while still allowing products from other intervals to be promoted if they have considerably high engagement scores. As a result, rigorous criteria for products to be highly content-relevant could potentially improve the ranking model's performance in content-based relevance.

2.3 Feature Generation

We employ a substantial number of features for ranking model training, with each feature being either engagement-related [1, 13, 16, 17, 20, 32, 39] or content-related [8, 23, 31]. For the content features, we use not only sparse features based on text match but also leverage LLMs to generate dense features. It is important to note a key difference in the size choice of language models for label vs. feature generation. Label generation can be performed completely offline where latency is not a concern, allowing us to use more complex LLMs to improve performance. However, features used in model training will also be computed at runtime during inference. While complex LLMs can deliver superior performance, their high computational cost presents a trade-off that must be carefully considered [28]. Therefore, we must ensure that the size of the LLMs used to generate content features is kept within an appropriate range to avoid latency degradation.

Leveraging our rich in-house query and product attribute information and expert judgment, we train a moderate-size BERT-based [6] model with the cross-encoder [9] framework to generate content-related features for ranking model training and inference.

3 Experiments and Evaluations

With the proposed design in Section 2, we trained seven models, including one Baseline model and six Variant models as candidates. In the Baseline model, instead of using LLMs, we employ an XGBoost [5] model trained with a few content features to generate content relevance scores for labeling. Additionally, the Baseline model does not include the cross-encoder (XE) features in training or inference. In the Variant models, we apply LLMs to generate labels and/or include the cross-encoder features for training and inference. Model details are listed in Table 1. In Variants L and LX, we directly use the LLM-predicted value as the content-based relevance score as part of the label. In Variants σ_cLX, σ_rLX, and σ_lLX, we apply 3 different sigmoid transformations listed below.

- **Variant σ_cLX** takes $\alpha = 12$ and $\beta = 0.5$, which only polarizes the high and low values without shifting the ***center***.
- **Variant σ_rLX** takes $\alpha = 10$ and $\beta = 0.7$, shifting the center to the ***right*** while keeping $\sigma(0)$ and $\sigma(1)$ close to 0 and 1, without introducing excessively steep gradients in any sub-intervals. This sets a higher threshold, i.e., a more rigorous criterion for high content relevance.
- **Variant σ_lLX** takes $\alpha = 10$ and $\beta = 0.3$, shifting the center to the ***left*** while keeping $\sigma(0)$ and $\sigma(1)$ close to 0 and 1, without introducing excessively steep gradients in any sub-intervals. This sets a lower threshold, i.e., a more relaxed criterion for high content relevance.

We plot the distribution of the original LLM-predicted scores and the transformed scores for all query-product pairs in the training data in Figure 2.

To measure the predictive performance of the Variant models, we evaluate both the content-based and engagement-based relevance of the ranking results generated by each model. The former is assessed through offline human judgment, while the latter is evaluated by conducting online interleaved A/B tests [4].

3.1 Offline Evaluation of Content Relevance

To evaluate the content-based ranking relevance of the Variant models, we performed an offline human evaluation for each of them using NDCG metrics. The process starts with sampling a substantial number of representative search queries across all segments. For

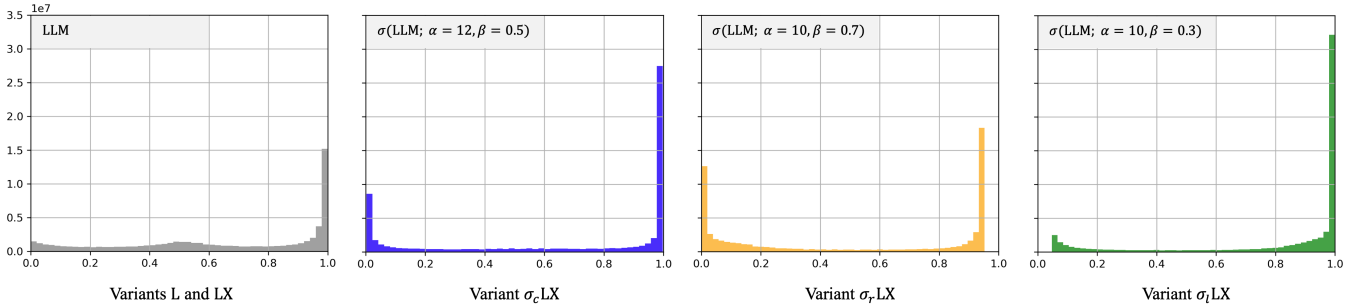

Figure 2: Distribution of LLM Scores and Transformed Scores in Histograms

Table 1: Baseline and Variant Models in Experiments

Model ID	Content Label	XE Features
Baseline	XGBoost	no
Variant X	XGBoost	yes
Variant L	LLM	no
Variant LX	LLM	yes
Variant σ_cLX	$\sigma(\text{LLM}; \alpha = 12, \beta = 0.5)$	yes
Variant σ_rLX	$\sigma(\text{LLM}; \alpha = 10, \beta = 0.7)$	yes
Variant σ_lLX	$\sigma(\text{LLM}; \alpha = 10, \beta = 0.3)$	yes

each query, we retrieve the top 10 products ranked by both the Baseline and Variant models. Human evaluators then assign 5-point relevance ratings to each query-product pair based on the product's relevance to the search query under well-defined guidelines. Finally, we calculate the NDCG@10 score for each query using the rating scores and item positions and conducted a T-test to compare the overall relevance quality between the Baseline and Variant models.

The evaluation results are listed in Table 2. The results lead to the following findings, which prove our hypotheses in Section 2.2.

- The comparison between Variants X, L, and LX suggests that simply adding XE features or incorporating LLMs in the label during model training alone does not result in a significant lift in ranking content relevance; however, combining these two approaches leads to a substantial improvement in content relevance.
- The comparison between Variants LX and σ_cLX indicates that the polarization used to push moderate relevances toward extremes during model training effectively enhances the model's predictive capability to prioritize highly content-relevant items in ranking.
- The comparison between Variants σ_lLX, σ_cLX, and σ_rLX suggests that, within an appropriate range, the more rigorous the criteria for high content relevance during model training, the greater the lift in content relevance achieved by the ranking model, and vice versa.

3.2 Online Tests for Engagement Measure

To measure the engagement performance of the Variant models, we conducted six interleaving tests, an online experiment where users are exposed to a mixed ranking list from both Baseline and Variant models on a substantial volume of online customer traffic at Walmart.com. The key metric is the percentage change in the number of items added to carts within the top 40 positions (ATC@40) for each Variant compared to the Baseline.

Table 2: Content Relevance Evaluation Results

Variant Model	NDCG@10 Change (p-value)
Variant X	+0.41% (0.11)
Variant L	+0.11% (0.60)
Variant LX	**+1.35**% (0.00)
Variant σ_cLX	**+1.72**% (0.01)
Variant σ_rLX	**+3.96**% (0.00)
Variant σ_lLX	−0.26% (0.48)

Table 3: Engagement Relevance Test Results

Variant Model	ATC@40 Change (p-value)
Variant X	−0.05% (0.68)
Variant L	−0.06% (0.60)
Variant LX	+0.05% (0.11)
Variant σ_cLX	+0.01% (0.91)
Variant σ_rLX	**−0.79**% (0.00)
Variant σ_lLX	**+1.04**% (0.00)

The interleaving test results are listed in Table 3. The results suggest that using XE features and/or LLM labels does not significantly affect ranking engagement. A significant impact is observed only when applying shifted sigmoid transformations to the LLM labels, where engagement performance changes in the opposite direction to content-based relevance.

Specifically, comparing the performance of Variants (σ_cLX vs. σ_rLX), (σ_cLX vs. σ_rLX), and (σ_rLX vs. σ_lLX), we observe a pattern where an increase in content relevance is accompanied by a compromise in engagement. This suggests a trade-off between content-based and engagement-based ranking relevances, as anticipated in Section 2.1. This observation highlights a pathway for search ranking model training: to construct a label incorporating both relevance aspects, apply appropriate transformations, and adjust parameter settings accordingly to achieve optimal performance toward specific relevance goals. As our selected candidates

Table 4: Feature Importance of Cross-Encoder LLM Feature

Model ID	Importance Rank	SHAP value
Variant X	8	0.1141
Variant LX	5	0.2645
Variant σ_cLX	3	0.3817
Variant σ_rLX	2	0.6381
Variant σ_lLX	7	0.1808

for balancing both relevances effectively, Variants LX and σ_cLX were further evaluated through comprehensive online A/B tests, and all business metrics, such as GMV and conversions, showed neutral results, validating that these models can achieve content relevance gains without compromising engagement.

3.3 Root Cause Analysis

To investigate the underlying causes of the observed evaluation/test results in depth, we establish a pathway by analyzing the relationship between feature importance, labels, and model performance. During model training, we calculate the averaged SHAP value [22] for each feature across all iterations and derive the ordered feature importance. Shown in Table 4 is the feature importance rank and SHAP value of the highest cross-encoder LLM feature in different Variant models.

We make four key noteworthy points here. Firstly, it is observed that utilizing LLM in labeling elevates the importance of the XE features in the model. This is because LLM enriches labels with more content-specific information, which better guides the content-related features to make more effective predictions. Consequently, the model relies more on these features during inference, resulting in more improved predictive capability to distinguish relevant from irrelevant products in ranking. Secondly, applying the sigmoid transformation further boosts the importance of the XE features due to its polarization effect. With more extreme content judgments embedded in label construction, the content-related features are able to make even more effective predictions, which further enhances the model's performance in determining content relevance. Thirdly, a more rigorous content criterion in labeling increases the importance of the XE features, while a more relaxed criterion reduces it. The reason is that greater rigor reduces the density of instances with high content relevance, making them more distinguishable, as shown in Figure 2. This allows the content features to function more effectively in discrimination. Lastly, as the importance of XE features increases, engagement features become relatively less important, potentially causing a decline in customer engagement during inference. This explains the trade-off between content relevance and engagement relevance in performance from a feature perspective.

4 Conclusion

In this paper, we propose a novel approach to product search ranking model training centered on the integration of Large Language Models (LLMs) for both label formulation and feature generation. By leveraging fine-tuned LLMs alongside user interaction data, we optimize the model's training objective to account for both content-based and engagement-based relevance. The use of LLMs in labeling also enhances the effectiveness of LLM-driven content features in the ranking model, promoting more content-relevant products to top positions. Additionally, applying appropriate transformations to LLM scores in the labels further refines the balance between content and engagement relevances. Comprehensive online and offline evaluations demonstrate that our approach can yield significant improvements in content relevance while maintaining decent user engagement. This work provides valuable insights for enhancing training objectives, optimizing relevance, and implementing LLMs in e-commerce product search ranking.

Acknowledgments

We would like to express our gratitude to Nguyen Vo and Changsung Kang from the Walmart Search Ranking Team for their assistance with the language model training that supported this study.

References

[1] E. Agichtein, E. Brill, and S. Dumais. 2019. Improving Web Search Ranking by Incorporating User Behavior Information. *SIGIR* (2019), 11–18. https://doi.org/10.1145/3308774.3308778

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*.

[3] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*. 129–136.

[4] O. Chapelle, Y. Zhang, and Y. Chang. 2012. Large-Scale Validation and Analysis of Interleaved Search Evaluation. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 233–242.

[5] T. Chen and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

[7] R. M. T. R. Eletreby, C. Mu, Z. Wang, and R. Mukherjee. 2022. Machine learning based methods and apparatus for automatically generating item rankings. US Patent App. 17/246,179.

[8] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. ACM, 2333–2338.

[9] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations (ICLR)*.

[10] K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[11] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L.R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06822* (2023).

[12] Z. Jiang, R. Tang, J. Xin, and J. Lin. 2021. How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

[13] Thorsten Joachims, Filip Radlinski, et al. 2017. User behavior modeling for ranking: A literature review. *Foundations and Trends in Information Retrieval* 9, 3 (2017), 151–255.

[14] J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations (ICLR)*.

[16] J. Liu, J. Zhao, J. Zheng, J. Zhao, C. Mu, and Z. Yan. 2024. Personalized search and browse ranking with customer brand affinity signal. US Patent App. 18/103,156.

[17] Q. Liu, A. Singh, J. Liu, C. Mu, Z. Yan, and J. Pedersen. 2024. Long or Short or Both? An Exploration on Lookback Time Windows of Behavioral Features

in Product Search Ranking. In *SIGIR eCom 2024*. https://sigir-ecom.github.io/eCom24Papers/paper_5.pdf

[18] S. Liu, L. Li, J. Song, Y. Yang, and X. Zeng. 2023. Multimodal Pre-Training with Self-Distillation for Product Understanding in E-Commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1039–1047. https://doi.org/10.1145/3539597.3570423

[19] T. Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[20] X. Liu, V. Joshi, H. JungWoo, C. Mu, and R. Mukherjee. 2024. Systems and methods for improving ecommerce search ranking via query-price affinity values. US Patent App. 17/308,477.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[22] S.M. Lundberg and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017), 4765–4774.

[23] A. Magnani, F. Liu, S. Chaidaroon, S. Yadav, P. Reddy Suram, A. Puthenputhussery, S. Chen, M. Xie, A. Kashi, T. Lee, and C. Liao. 2022. Semantic Retrieval at Walmart. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 3495–3503.

[24] R. Nogueira and K. Cho. 2019. Passage Re-ranking with BERT. In *arXiv preprint arXiv:1901.04085*.

[25] R. Nogueira, W. Yang, K. Cho, and J. Lin. 2019. Multi-stage Document Ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

[26] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[27] OpenAI. 2023. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2304.03442* (2023).

[28] R. Pradeep, R. Nogueira, and J. Lin. 2021. On the Effectiveness of Small Language Models for Query and Document Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1056–1060.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[30] S. K. K. Santu, P. Sondhi, and C. Zhai. 2017. On application of learning to rank for e-commerce search. In *Proceedings of SIGIR*.

[31] H. Shan, Q. Zhang, Z. Liu, G. Zhang, and C. Li. 2023. Beyond Two-Tower: Attribute Guided Representation Learning for Candidate Retrieval. In *Proceedings of the ACM Web Conference 2023*. ACM, 3173–3181.

[32] L. Sun, S. Zhao, and Q. Yang. 2022. Personalized Search Ranking Based on User Engagement Behaviors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1234–1243.

[33] Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295* (2024).

[34] P. Thomas, S. Spielman, N. Craswell, and B. Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1930–1940. https://doi.org/10.1145/3626772.3657707

[35] W. Tong, J. Liu, J. Zhao, N. Baliga, and Z. Yan. 2022. Event-based Learning to Rank Framework to Personalize Search Ranking. United States patent application, filed Jan. 30, 2022.

[36] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Roziere, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, D. Bashlykov, and S. Batra. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[38] N. Vo, H. Shang, Z. Yang, J. Lin, S.D. Mohseni Taheri, and C. Kang. 2024. Knowledge Distillation for Efficient and Effective Relevance Search on E-commerce. In *SIGIR eCom 2024*. https://sigir-ecom.github.io/eCom24Papers/paper_16.pdf

[39] Y. Wang, D. Zhu, A. P. Ruchandani, C. Mu, Y. M. Sun, and S. Agarwal. 2024. Systems and methods for facilitating online search based on offline transactions. US Patent App. 17/389,111.

[40] Y. Yang, Y. Qiao, J. Shao, X. Yan, and T. Yang. 2022. Lightweight composite re-ranking for efficient keyword search with BERT. In *Proceedings of WSDM*.

[41] S. Yao, J. Tan, X. Chen, J. Zhang, X. Zeng, and K. Yang. 2022. ReprBERT: Distilling BERT to an Efficient Representation-Based Relevance Model for E-Commerce. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022), 4363–4371. https://doi.org/10.1145/3534678.3539090

[42] A. Yates, R. Nogueira, and J. Lin. 2021. Pre-trained Transformers for Text Ranking: BERT and Beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*. 1–4.