LLM-as-a-Judge in Entity Retrieval: Assessing Explicit and Implicit Relevance

Anonymous Author(s)

ABSTRACT

Entity retrieval plays a critical role in information access systems, yet the development and evaluation of retrieval models remain constrained by the limited availability of high-quality supervision. While recent work has demonstrated the utility of large language models (LLMs) as relevance assessors in passage and document retrieval, their reliability in the context of entity retrieval—where targets are abstract, underspecified, and often semantically sparse—remains unexplored. In this work, we evaluate LLM-based judgments against two complementary supervision signals: human-annotated relevance labels from the DBpedia-Entity benchmark and implicit feedback from user clicks in the LaQuE dataset. We show that LLMs exhibit strong agreement with expert annotations and replicate user click patterns with over 91% agreement, suggesting alignment with behavioral judgments despite noisy input queries. We further identify and analyze systematic mismatches for user clicks on irrelevant entities. Our findings establish LLMs not only as effective annotators for entity relevance judgment—even when given only the entity title—but also as powerful tools for predicting click-through behavior and simulating explainable user intent. Our code, prompts, and data are publicly available at: https://anonymous.4open.science/r/ClickLLM-E812

1 INTRODUCTION

Entity retrieval is central to modern information access, underpinning web search, digital assistants, and academic platforms. Studies show that a large portion of queries—over 70% on Bing and more than half on Semantic Scholar—are entity-focused [7, 21, 35]. This growing demand is supported by knowledge graphs like DBpedia, Wikidata, and YAGO [15, 30, 34], which provide structured representations of entities and serve as foundational resources for both retrieval systems and language model training [24, 25]. Despite this importance, progress in entity retrieval is constrained by the scarcity of high-quality evaluation benchmarks [5, 6, 9]. In contrast to ad hoc passage retrieval, where datasets such as MS MARCO have enabled the development of data-intensive neural systems, entity retrieval suffers from a lack of diverse, large-scale, and manually curated datasets [23].

Existing resources such as the DBpedia-Entity collections remain limited in scale and do not support robust training or reliable fine-grained evaluation [16]. To address this gap, recent work has turned to silver-standard resources such as LaQuE, which uses click-through data from the ORCAS dataset to construct over 2 million query-entity relevance pairs [2, 10]. Despite offering scale and behavioral diversity, these annotations are based on implicit feedback from user clicks rather than expert judgments, limiting their reliability [19, 20]. In this work, we investigate the potential of large language models (LLMs) to serve as relevance assessors for entity retrieval. While LLMs have demonstrated strong alignment with human judgments in document and passage retrieval [1, 12, 22, 31, 33], their applicability to entity-centric tasks where targets are abstract, structured, and often underspecified remains uncertain [29]. Our study evaluates whether LLM-based judgments can complement or reinforce both expert annotations and click-derived feedback. Unlike prior work [14, 36],which focuses on improving LLM efficiency for long user behavior sequences in click through rate prediction, our work investigates the use of LLMs for interpretable relevance assessment and user behavior analysis in entity-centric search, where user queries target real-world entities.

Our paper offers the following **contributions**: **(1)** We provide the the first systematic analysis of LLM-based relevance judgments in entity retrieval, comparing them against both expert labels (DBpedia-Entity) and click-derived labels (LaQuE). **(2)** We examine the alignment between LLM judgments and user clicks, identifying where the two agree and where they diverge, with a focus on understanding whether LLMs can recover semantic relevance from behavioral signals. **(3)** We introduce a novel analysis of click-relevance mismatches by isolating cases where users clicked on entities unanimously judged irrelevant. We generate plausible explanations for these behaviors, revealing common patterns such as lexical confusion and familiarity-driven exploration.

Our findings show that LLMs can produce relevance judgments for entities, which align closely with both expert annotations and user click behavior. On the DBpedia-Entity benchmark, LLMs achieve agreement scores comparable to those reported for passage retrieval tasks. In LaQuE dataset, LLMs correctly identify over 91% of relevant clicked entities, demonstrating high alignment with implicit user feedback. Notably, when user clicks diverge from LLM-based relevance judgments, our analysis shows that these disagreements are not random but follow consistent behavioral patterns. By systematically identifying and categorizing these patterns, LLMs are able to do beyond relevance judgment by explaining why users behave the way they do, offering insights into common sources of noise or misunderstanding in click-based data during entity search.

2 OVERVIEW OF METHODOLOGY

This study is motivated by two objectives **(1)** evaluating whether LLMs can function as high-quality relevance assessors for entity retrieval, and **(2)** uncovering how their judgments relate to both expert labels and behavioral signals derived from user clicks. Entity retrieval presents distinct challenges compared to passage or document retrieval: entities are often short, semantically abstract, and tied to structured representations [35], while user queries seeking entities are often underspecified and multifaceted [11, 28]. These characteristics complicate both annotation and evaluation of relevant entities for user queries. Traditional supervised resources for this task are limited in scale and coverage, and behavioral data such as clicks are abundant but noisy. Our approach aims to bridge this gap by positioning LLMs as scalable, interpretable, and semantically grounded judgment agents capable of operating in both high- and low-supervision environments[8, 26, 27]. This aligns directly with

Table 1: Example queries and relevance judgments from DBpedia-Entity.

Query	Entity	Relevance Judgment	
		LLM	Human
Einstein Relativity theory	`Theory of Relativity`	2	2
Disney Orlando	`Greater Orlando`	1	0
Austin Texas	`Texas`	0	1
Guitar Classical Bach	`Johann Sebastian Bach`	2	0

Table 2: Example queries from the LaQuE dataset with corresponding clicked entities and LLM relevance judgments.

Query	Clicked Entity	LLM Judgment
Apple Mac	`Macintosh`	Relevant
Indian History in Hindi	`Hindi`	Not Relevant
CNN News Cast Members	`List of CNN Anchors`	Relevant
When Was Color Invented	`Color television`	Not Relevant

our core research questions concerning the alignment of LLM judgments with expert annotations, their agreement with user behavior, and their capacity to explain divergences between the two.

We adopt a three-stage evaluation framework. **First**, we assess LLM relevance judgments against expert-labeled data from DBpedia-Entity v2. We prompt two LLMs, Qwen and LLaMA, under two input conditions: `query + entity title` and `query + entity title + abstract`. These variations allow us to measure the relative influence of surface cues and contextual grounding. Relevance is evaluated in both binary and graded formats [3, 12, 33], enabling comparisons across granularities. Agreement with human labels is quantified using Cohen's kappa, and confusion matrices are analyzed to identify systematic judgment patterns. **Second**, we evaluate LLM judgments against behavioral feedback from LaQuE, a large-scale dataset derived from real-world user clicks [2]. As LaQuE lacks negative labels, we treat clicked entities as implicitly relevant and compute click-agreement under both input conditions. This helps assess whether LLMs can recover semantic relevance from implicit feedback. While this evaluation lacks the symmetry of expert-labeled datasets, it offers scale and diversity due to the data size. **Third**, to understand sources of disagreement between LLMs and user click behavior, we isolate a filtered set of query-entity pairs where users clicked on entities consistently judged as irrelevant. We then use LLMs to generate plausible explanations for these mismatches, hypothesizing why a user might click on an entity despite its irrelevance. This provides diagnostic insight into the epistemic and behavioral divergences between relevance and observed user behavior.

3 EMPIRICAL SETUP

3.1 Datasets

We evaluate LLM-based relevance judgments using two benchmark datasets for entity retrieval: DBpedia-Entity v2 [16] and LaQuE [2]. **DBpedia-Entity v2** is one of the most widely used resources for evaluating entity retrieval models. It contains 485 queries spanning four types, namely entities, keyword queries, entity lists, and natural language questions. Each query is annotated with graded human relevance labels on a 3-point scale.In total, the dataset includes over 50,000 query-entity relevance judgments. **LaQuE**, in contrast, is a large-scale resource constructed from clickthrough logs in the ORCAS dataset [10]. It consists of over two million real-user queries paired with Wikipedia entities that users clicked on, enabling large-scale training and evaluation in realistic settings. For this study, we randomly sample 15,000 queries from LaQuE, yielding 16,218 query-entity pairs. Each pair corresponds to an entity clicked by a user in response to the query. Unlike DBpedia, LaQuE provides only positive click signals, i.e., it lacks explicit non-relevance labels for unclicked entities. While DBpedia supports fine-grained evaluation with full graded supervision, LaQuE allows us to assess whether LLMs can approximate user behavior and identify relevant entities from implicit feedback. Together, these datasets enable a comprehensive examination of LLM-based judgments across both explicitly annotated and behavior-driven relevance signals.

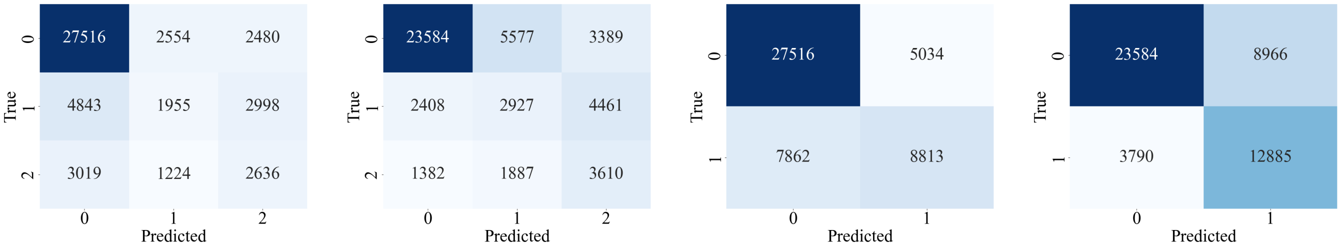
3.2 LLM-based relevance judgment

Our goal is to evaluate whether LLMs can act as reliable judges across expert-labeled and behavior-derived signals and under varying levels of contextual input for entity search. We conduct experiments using two LLMs: Qwen3:8b and LLaMA4:Scout. Both are evaluated under two supervision settings (expert annotation and click data) and two input conditions (query + entity title to explore whether the model's internal knowledge is sufficient to determine relevance, and query + entity title + abstract to resolve any possible lexical ambiguity). This design allows us to assess the models' ability to reason about entity relevance based on prior knowledge alone, as well as with structured contextual support. Due to space constraints, we report only results for Qwen in the main text; outcomes for LLaMA, which follow similar trends, are available on our GitHub repository. While prior work has commonly used the UMBRELLA prompt [31, 33] for eliciting graded relevance scores, we found that it generalizes poorly to entity retrieval. Entities tend to be short, lack lexical continuity, and depend more heavily on factual grounding than topical elaboration. To address this, we adapt the UMBRELLA framework by explicitly guiding the model through three reasoning steps: (1) estimating the likely user intent behind the query, (2) measuring factual or conceptual alignment between the query and entity, and (3) producing a final judgment. This modified prompt is is made publicly available for replication on our Github repository. For DBpedia-Entity, we evaluate model outputs against human annotations using a three-grade scale (0 = irrelevant, 1 = relevant, 2 = highly relevant), reporting both graded and binarized agreement metrics. For LaQuE, we treat clicked entities as implicitly relevant and measure whether the LLMs assign relevance labels consistent with user behavior. Representative examples from each dataset are shown in Table 1 and Table 2.

4 FINDINGS

4.1 Agreement with Human Annotation

Our goal in this analysis is to evaluate how closely LLM-based relevance judgments align with expert human annotations in the DBpedia-Entity v2 dataset. We consider both graded and binary relevance settings and assess performance under two input conditions: using the entity title alone and using both the title and DBpedia abstract. This setup allows us to examine the role of contextual information in supporting entity-level judgments, which are typically more abstract and structurally sparse than document-based tasks.



(a) Title only (Graded) **(b) Title + Abstract (Graded)** **(c) Title only (Binary)** **(d) Title + Abstract (Binary)**

Figure 1: Comparing LLM-based relevance judgments to human annotations on DBpedia-Entity.

Table 3: Cohen κ scores for the binary and graded LLM judgement from DBpedia-Entity

Input	Binary	Graded
Titles	0.3900	0.2733
Titles + Abstracts	0.4623	0.3042

Table 4: Agreement between LLM relevance judgments and Click-through data on 15k queries.

Input	#Agreements	Accuracy
Titles	14, 910	91.93%
Titles + Abstracts	14, 888	91.79%

Figure 1 presents the confusion matrices for all configurations. In the graded setting (Figures 1a and 1b), where relevance is assigned on a 3-point scale (0 = irrelevant, 1 = relevant, 2 = highly relevant), the model shows strong performance on identifying irrelevant entities when using titles alone, correctly classifying 27,516 instances. However, its ability to identify highly relevant entities is limited, with only 2,636 level-2 entities labeled correctly. When abstracts are included, the number of correctly classified level-2 entities increases to 3,610, a 37% improvement. This gain in recall comes at a cost, namely correctly identified irrelevant entities drop to 23,584, and false positives at level 2 increase from 2,480 to 3,389.

This trade-off illustrates a consistent pattern where *adding contextual information improves recall for relevant content but introduces noise for non-relevant cases.* The inclusion of the abstract also appears to help sharpen distinctions between grade 1 and grade 2. In the title-only setup, the model tends to overuse grade 1, likely due to insufficient disambiguation, whereas the abstract enables more confident and accurate assignments to grade 2. This is evidenced by the increase from 2,636 to 3,610 correct grade-2 predictions.

In the binary condition (Figures 1c and 1d), relevance labels are collapsed into 0 (irrelevant) and 1 (relevant). With the title-only input, the model correctly identifies 27,516 irrelevant and 8,813 relevant entities. When abstracts are added, the number of correct relevant classifications increases to 12,885, a 46% improvement. Interestingly, this improvement in recall is accompanied by a reduction in false positives where irrelevant entities misclassified as relevant decrease from 7,862 to 3,790. This suggests that in some cases, abstract information helps the model reject entities that exhibit surface-level lexical similarity to the query but are topically unrelated. These patterns indicate that titles alone serve as effective coarse-grained signals for filtering irrelevant entities, while abstracts provide the necessary disambiguation to recover higher-relevance items especially in ambiguous or underspecified cases. However, the added context can also blur boundaries between relevance classes, leading to increased confusion in borderline instances. This precision–recall trade-off is especially important in high-precision applications [13, 17].

To quantify agreement with expert annotations, Table 3 reports Cohen's Kappa coefficients for both input conditions. With titles only, the binary setting yields $\kappa = 0.3900$ and the graded setting $\kappa = 0.2733$. When abstracts are added, these values increase to $\kappa = 0.4623$ and $\kappa = 0.3042$, respectively. These levels of agreement fall within the range reported in recent work on LLM-based relevance judgment for document and passage retrieval, including evaluations using the UMBRELLA framework on TREC DL 2019–2023 benchmarks []. Given the increased sparsity, ambiguity, and lack of lexical overlap in entity retrieval [], these results are encouraging. They suggest that LLMs can approximate expert-level relevance decisions even with minimal input, and that structured contextual information further improves alignment.

4.2 Agreement with Click-through Data

We also examine how well LLMs align with user click behavior i.e., we study the agreement between the LLM's predictions and the entities that users actually clicked on in the LaQuE dataset. Unlike DBpedia-Entity, the LaQuE dataset consists of real-world user queries, often short, noisy, and loosely structured. These queries are not professionally curated and exhibit considerable variation in spelling, grammar, and intent, which makes relevance estimation more challenging. We conduct the same experimental setup as in the previous section: both binary and graded relevance judgments are generated using two LLM input settings, namely (1) query + entity title, and (2) query + entity title and abstract. However, a key difference is that LaQuE only provides information about entities that were clicked. There are no explicit non-relevant judgments for unclicked entities. As a result, we cannot compute confusion matrices or standard agreement metrics such as Cohen's Kappa. Instead, we report the proportion of clicked entities that were labeled as relevant or highly relevant by the LLMs under each condition. This allows us to assess how well the LLMs align with implicit user judgments, albeit from a one-sided (positive-only) perspective. Table 4 reports the agreement between LLM-based relevance judgments and user click behavior on the LaQuE dataset. Despite the absence of explicit relevance labels, LLMs show strong alignment with implicit user preferences. Using only the entity title, the model agrees with user clicks in 91.93% of cases; adding the entity abstract yields a similar rate of 91.79%, suggesting that even minimal input suffices to replicate user behavior at scale. While abstracts do not

Table 5: Example queries from the LaQuE dataset with corresponding clicked entities and LLM click-through reasoning.

Query	Clicked Entity	LLM Judgement
LLM Reasoning		
X Factor USA Judges	`The X Factor (UK TV series)`	Name or lexical similarity, Prominent Results Bias
Palm Springs Florida	`Palm Springs, California`	Name or lexical similarity, Geographic name confusion
Brad Pitt Vegan	`List of vegans`	Category or topical association, Prominent Result Bias, Exploratory curiosity
John Bundy	`Ted Bundy`	Name or lexical similarity, Prominent Result Bias, Exploratory curiosity, Familiarity/Recognition Bias

notably improve overall agreement, they increase the proportion of entities judged as highly relevant, indicating LLMs can capture more nuanced signals with additional context. The slight drop in agreement may reflect the model's ability to reject entities clicked due to superficial cues like lexical overlap or position bias. These findings highlight the potential of LLMs to enhance click through data and extract meaningful relevance signals from noisy feedback.

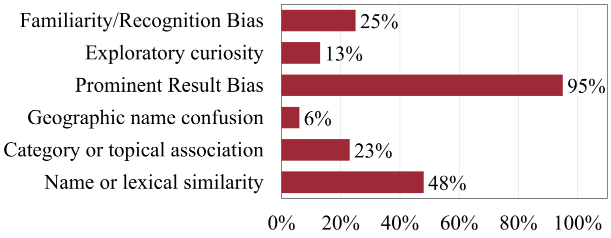


Figure 2: Distribution of LLM-generated reasons for user clicks on entities judged irrelevant. Prominent result bias and lexical similarity are the most frequent factors.

4.3 User Clicks on Non-Relevant Entities

User clicks are often treated as implicit indicators of relevance, yet they do not always reflect genuine semantic alignment between query and result. Instead, clicks may arise from misunderstanding, ambiguity, interface design, or user intent that diverges from the literal query. The goal of this analysis is to investigate such mismatches, specifically cases where users clicked on entities that were consistently judged irrelevant in order to identify systematic patterns in user behavior that lead to misleading feedback signals in silver standard datasets such as LaQue. To ensure that we are analyzing truly irrelevant clicks and not cases where the LLM assessments may themselves be noisy or uncertain, we adopt a conservative filtering strategy. For each query-entity pair in the LaQuE dataset, we obtain relevance judgments from four independent LLM configurations: two models (LLaMA and Qwen) and two input settings (query + entity title, query + entity title and abstract). We retain only those query-entity pairs labeled as irrelevant by at least three of the four configurations. This consensus-based filtering ensures high confidence in the irrelevance assessments and enables us to focus on well-substantiated disagreements between user behavior and model judgments. Using this approach, we identify 420 query-entity pairs that were clicked by users despite being consistently rated as irrelevant by the models.

To understand why users might have clicked on these entities, inspired by [27] we adopt an *interpretable autoprompting* strategy. This technique facilitates the extraction of human-interpretable rationales from LLMs, aiding in the interpretation of complex user behaviors. To this end, we adopt this approach and likewise prompt an LLM (Qwen) with the query and the clicked (but judged-irrelevant) entity. The prompt clearly states that the entity is not relevant and asks the model to hypothesize plausible user motivations for the click. These explanations are grounded in known patterns of search behavior, such as attention to surface features, curiosity, or cognitive heuristics [4, 18, 32]. To avoid redundancy and overfitting to individual cases, we then prompt to summertime the list of reasons over user's motivation to click on irrelevant entities. The final set of categories captures diverse sources of mismatch, including but not limited to 'lexical similarity', 'ambiguity', 'layout bias', and 'user familiarity' and more. After obtaining these categories, we do reason assignment i.e., assigning one or more reasons to each query-entity pair through the LLM. Table 5 presents representative examples of such cases, including the original query, the clicked entity, and the rationale generated by the LLM. Figure 2 summarizes the frequency of these reasons across the filtered dataset. The most common explanation was `Prominent Result Bias`, suggesting that users are strongly influenced by interface placement and visual salience consistent with well-established findings in click behavior research [18]. The next most frequent category was `Lexical Overlap` where users clicked on entities that matched query terms on the surface level but were semantically off-topic. Other notable reasons included `Familiarity or Recognition`, where users clicked based on prior knowledge rather than relevance, and `Category-Level Association`, reflecting loose conceptual ties that fall short of strict topical relevance. While these reasons have not been validated through human annotation, they offer a promising first step toward more explainable and interpretable evaluation of entity-centric retrieval behavior.

This analysis reveals that LLMs are not only useful as scalable relevance assessors but also as diagnostic tools for interpreting noisy user behavioral data when performing entity search. By identifying the latent factors that drive user clicks, even when misaligned with query intent, LLMs can help disentangle relevance from attention and improve the interpretability of implicit feedback in entity-centric retrieval settings.

5 CONCLUSION

This work demonstrates that large language models (LLMs) are effective tools for both evaluating and interpreting relevance in entity retrieval. On both expert-labeled (DBpedia-Entity) and behavior-based (LaQuE) datasets, LLMs align well with ground-truth and user click data, especially when provided with contextual signals like entity abstracts. Beyond assessment, LLMs can generate plausible explanations for user clicks on irrelevant entities, revealing systematic patterns such as interface bias, lexical overlap, and familiarity effects. These capabilities suggest that future evaluation frameworks can integrate LLM-based diagnostics to better interpret noisy user behavior—particularly in domains where labeled data is limited and click signals are imperfect.

GENAI USAGE DISCLOSURE

The authors affirm that generative AI tools were not used in the research design, data collection, analysis, or in generating substantive content. These tools were used solely for minor editing and grammar refinement. All technical contributions and written material were primarily created and thoroughly verified by the authors.

REFERENCES

[1] Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*. Springer, 135–159.

[2] Negar Arabzadeh, Amin Bigdeli, and Ebrahim Bagheri. 2024. Laque: Enabling entity search at scale. In *European Conference on Information Retrieval*. Springer, 270–285.

[3] Negar Arabzadeh and Charles LA Clarke. 2025. Benchmarking LLM-based Relevance Judgment Methods. *arXiv preprint arXiv:2504.12558* (2025).

[4] Anne Aula, Rehan M Khan, and Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 35–44.

[5] Krisztian Balog. 2018. Entity Retrieval.

[6] Krisztian Balog and Robert Neumayer. 2013. A test collection for entity search in DBpedia. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 737–740.

[7] Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. 2010. *Overview of the TREC 2010 entity track*. Technical Report. NORWEGIAN UNIV OF SCIENCE AND TECHNOLOGY TRONDHEIM.

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.

[9] Shubham Chatterjee, Zhen Zhang, Shraman Ghosh, Francis Ferraro, Bhaskar Mitra, Paul McNamee, Tim Finin, and James Mayfield. 2023. ELI5Entity: A Unified Benchmark for Explainable Entity Linking and Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 344–355.

[10] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. Orcas: 18 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2983–2989.

[11] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity Query Feature Expansion using Knowledge Base Links. In *Proceedings of the 37th international ACM SIGIR conference on Research development in information retrieval*. ACM, 365–374.

[12] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi (Eds.). ACM, 39–50. https://doi.org/10.1145/3578337.3605136

[13] Yue Feng, Fattane Zarrinkalam, Ebrahim Bagheri, Hossein Fani, and Feras Al-Obeidat. 2018. Entity linking of tweets based on dominant entity candidates. *Social Network Analysis and Mining* 8 (2018), 1–16.

[14] Binzong Geng, Zhaoxin Huan, Xiaolu Zhang, Yong He, Liang Zhang, Fajie Yuan, Jun Zhou, and Linjian Mo. 2024. Breaking the length barrier: Llm-enhanced CTR prediction in long textual user behaviors. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2311–2315.

[15] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting entity linking in queries for entity retrieval. In *Proceedings of the 2016 acm international conference on the theory of information retrieval*. 209–218.

[16] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. DBpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1265–1268.

[17] Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving Supporting Evidence for Generative Question Answering. *arXiv preprint arXiv:2309.11392* (2023).

[18] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR ’05)*. Association for Computing Machinery, New York, NY, USA, 154–161. https://doi.org/10.1145/1076034.1076063

[19] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007), 7–es.

[20] Seikyung Jung, Jonathan L Herlocker, and Janet Webster. 2007. Click data as implicit relevance feedback in web search. *Information processing & management* 43, 3 (2007), 791–807.

[21] Xinshi Lin, Wai Lam, and Kwun Ping Lai. 2018. Entity retrieval in the knowledge graph with hierarchical entity type and content. In *Proceedings of the 2018 acm sigir international conference on theory of information retrieval*. 211–214.

[22] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *SIGIR*. 2230–2235.

[23] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *choice* 2640 (2016), 660.

[24] Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. 2023. Large language models and knowledge graphs: Opportunities and challenges. *arXiv preprint arXiv:2308.06374* (2023).

[25] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599.

[26] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track. In *European Conference on Information Retrieval*. Springer, 132–148.

[27] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework. *arXiv preprint arXiv:2411.09607* (2024).

[28] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2016. Entity Aspect Linking: Providing Fine-Grained Semantics of Entities in Context. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1201–1210.

[29] Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 23–31.

[30] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics* 6, 3 (2008), 203–217.

[31] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences, 2023. *URL https://arxiv.org/abs/2309.10621* (2023).

[32] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.

[33] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *arXiv preprint arXiv:2406.06519* (2024).

[34] Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, et al. 2020. Wikidata as a knowledge graph for the life sciences. *Elife* 9 (2020), e52614.

[35] Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW ’17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1271–1279. https://doi.org/10.1145/3038912.3052558

[36] Huachi Zhou, Wenqi Pei, Qinggang Zhang, Daochen Zha, Hao Chen, and Xiao Huang. [n. d.]. Self-Monitoring Large Language Models for Click-Through Rate Prediction. ([n. d.]).