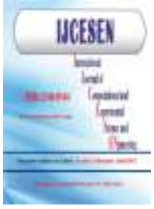
International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 12-No.1 (2026) pp. 439-446
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Leveraging Large Language Models for Search Relevance Measurement in E-Commerce

Prathyusha Bhaskar Karnam*

Independent Researcher, USA
* **Corresponding Author Email:** prathyusha.knm@gmail.com - **ORCID:** 0000-0002-5887-7996

Article Info:

DOI: 10.22399/ijcesen.4818
Received : 29 November 2025
Revised : 10 January 2026
Accepted : 20 January 2026

Keywords

Search Relevance Measurement,
Large Language Models,
E-Commerce Search Quality,
Chain-Of-Thought Prompting,
Automated Relevance Assessment

Abstract:

Search relevance measurement represents a critical challenge in e-commerce systems, where the accuracy of query-item matching directly impacts user experience, conversion rates, and platform trust. Traditional approaches to measuring relevance have relied heavily on human annotation and behavioral signals, both of which present significant limitations in scalability, cost, and accuracy. Human evaluation suffers from inter-rater variability, limited coverage of long-tail queries, and prohibitive resource requirements when applied to large product catalogs. Implicit feedback signals such as clicks and conversions introduce noise through position bias and popularity effects, often failing to reflect true relevance. Recent advances in large language models offer a transformative alternative by leveraging semantic understanding, contextual reasoning, and world knowledge to assess query-item relationships at scale. Through careful prompt engineering, chain-of-thought reasoning, and validation against human-labeled datasets, these models can generate reliable relevance judgments that approximate or exceed human performance while covering vastly larger evaluation spaces. Implementation strategies, including teacher-student architectures, active learning for edge cases, and periodic human audits, enable organizations to balance accuracy with operational efficiency. This article addresses fundamental trade-offs between coverage, quality, and cost that have long constrained traditional relevance measurement methodologies, enabling more sophisticated search systems that genuinely understand and respond to user intent rather than merely optimizing for engagement metrics.

1. Introduction

Search functionality serves as the critical bridge between user intent and product discovery in digital commerce platforms. The quality of this interaction hinges on relevance—the degree to which returned results align with what users actually seek. When search results accurately reflect user needs, the outcome is improved conversion rates, reduced friction, and strengthened platform trust. Research has demonstrated that search relevance significantly impacts business outcomes, with studies showing that leveraging contextual information and entity relationships in search systems can substantially enhance user experience and engagement metrics. The challenge lies in developing methods that can accurately capture and measure this relevance across diverse user queries and vast product catalogs [1]. However, measuring relevance presents fundamental challenges because it is neither directly observable nor objective. A user's perception of relevance depends on contextual factors, including their immediate needs, prior knowledge, and personal preferences, making it inherently subjective and difficult to quantify through direct observation.

Traditional approaches relying on manual annotation or behavioral proxies have proven expensive and difficult to scale. Human annotation campaigns typically require trained evaluators to assess thousands of query-item pairs, with each annotation demanding careful consideration and consistent application of relevance criteria. The challenge of achieving consensus among human judges represents a persistent problem in information retrieval evaluation. Research examining the exchangeability of relevance judges has revealed significant variability in how different assessors interpret and apply relevance criteria, even when provided with identical guidelines and training materials. Studies have found that while judges can be considered partially exchangeable in their assessments, the degree of agreement varies considerably across different types of queries and result sets. This variability introduces noise into training datasets and evaluation metrics, limiting the reliability of downstream applications [2]. The inconsistency becomes particularly pronounced when evaluating nuanced distinctions between categories such as exact matches versus suitable substitutes, or when determining whether complementary products constitute relevant results for a given query.

Furthermore, as product catalogs expand to millions of items and query distributions evolve with seasonal trends and emerging user needs, maintaining comprehensive coverage through manual annotation becomes increasingly impractical. The cost and time investment required to label even a representative sample of query-item pairs across a large catalog can be prohibitive. Organizations must balance the need for high-quality ground truth data against practical constraints of budget and timeline, often resulting in sparse coverage that fails to adequately represent long-tail queries or newly introduced products. The dynamic nature of digital commerce, where inventory changes frequently, and user search patterns shift with trends and seasons, further compounds these challenges by requiring continuous relabeling efforts to maintain dataset currency.

Recent advances in large language models present a promising alternative, offering scalable, semantically sophisticated relevance assessment capabilities that can transform how search systems evaluate and improve their performance. These models bring capabilities that extend beyond simple keyword matching, including understanding semantic relationships between queries and products, reasoning about context and user intent, and leveraging broad knowledge to evaluate relevance in ways that approximate human judgment while operating at significantly greater scale and speed. By automating the relevance assessment process, large language models can process vast quantities of query-item pairs efficiently, providing consistent evaluations based on learned patterns and contextual understanding that can match or exceed human-level performance on many relevance judgment tasks.

2. The Challenge of Implicit Signals

Relevance is usually determined by user behavior, like clicks, adding to cart, and purchases, which are the features of search systems. Although these signals are useful to get feedback, they bring a significant amount of noise to relevance assessment. Implicit feedback mechanisms also pose inherent problems in the accurate measurement of the quality of search since the interaction of the user can be influenced by other factors other than the actual relevance. Studies have established that click-through activity, despite its frequent application as a relevance measure, is strongly a presentation factor in terms of result position, quality of thumbnails, and pricing information as opposed to actual item query alignment. Research investigating the correlation between user clicks and relevance in reality has found that there have been significant discrepancies in this, with position bias being the most significant. Experimental data prove that users tend to click on the results posted higher in the list more often when the results of lower ranks are more relevant to their queries. This position bias can be so great that documents of no importance that are placed on top of the search results get many more clicks than the highly relevant documents that are ranked lower. The study suggests that it is important to consider such biases when analyzing clickthrough data because the crude click information cannot be directly converted into relevance ratings and can lead to false alarms in optimizing a given system without necessary adjustments [3]. Remarkable or trending objects can also win interest even though they are not meeting the actual goal of the user, and end up in the case where search systems that are trained to maximize the engagement metrics may inadvertently promote the attractive and therefore less relevant elements rather than the ones that suit the user best.

Alternatively, less relevant items that are still of high value and importance, and those that are not presented with a strong appeal to search engines, can get little or no interaction, which forms a vicious circle of less-than-ideal rankings. This is especially harmful to novice products that do not have historical engagement data or deep tail queries, where the user might give up their search without scrolling to the bottom of the page. Coupled with the lack of observability of user behavior and actual relevance, it is hard to understand what is actually wrong with quality, how a ranking algorithm can be successfully trained, or whether the improvement in engagement is due to an actual increase in relevance or to the popularity effect. When models are trained mostly using click data, there is a danger that they will learn to predict popularity, at the expense of relevance, which is worse than the search experience of a user whose needs are not typical in the mainstream.Structured relevance taxonomies overcome this disadvantage by changing relevance into a measurable object. Quintessential frameworks are used to classify query-item relations into specific groups: an exact match, which perfectly satisfies the query; substitutes, which provide an alternative similar in functionality; complements, which are naturally paired with the item sought; and irrelevant responses. Such taxonomies have found more and more adoption in the e-commerce search evaluation, and frameworks that offer standardized schemas to evaluate query-product relationships have been proposed. The new studies on product search and recommendation systems have stressed the need to ensure development of methodologies that are effective in search relevance evaluation in business situations. Research has indicated that explicit relevance assessment frameworks are more efficient in the measurement of search system performance as opposed to the measurement of search system performance only based on behavioural indicators especially in contexts where user intent is multifaceted or complex [4]. These explicit labels may be used in a variety of downstream tasks, such as to supervise training of ranking models, to provide rule-based ordering logic with information, and to make a precise evaluation metric, like the percentage of exact matches on the first page or the frequency of irrelevant items on the first page.

Such a systematic methodology provides search groups with the ground truth that would subsequently be useful in systematic quality evaluation and refinement. The organizations can overcome the shortcomings of implicit feedback by defining each category of relevance and creating labeled datasets to reflect these differences, and creating s, search systems that can truly comprehend and act on user intent.

3. Traditional Relevance Measurement Approaches

Measurement of relevance based on history. The analysis of relevance based on history has been focused on human assessment of query-item pairs based on predetermined procedures. Each item is evaluated by annotators in the context of the inferred user intent to which they assign the relevant labels of their detailed taxonomies. This method involves a significant investment in the training of annotators so as to have consistency and reliability in the judgments. Studies that investigate human relevance assessment have shown that, although trained judges are able to reach a reasonable consensus on cases where there is an apparent cutoff, the judgments of the judges are not always consistent on marginal cases, where the connection between query and item entails subtle semantic differences. However, research studies have demonstrated that guidelines being applied must be clear, the annotators must be experts, and the taxonomy under consideration must be complex or not. Intense work on the construction of strong evaluation strategies has especially played a significant role in the field of information retrieval research, where the demand for strong test collections has prompted new advancements in assessment design and quality management. The studies of evaluation frameworks state that relevant evaluation judgment cannot be done with a clear set of criteria; one must take into account the context of the user, the purpose of the task, and the dynamism of the information requirements. The introduction of standardized methods of developing and testing relevance tests has made it possible to perform more rigorous experimental analysis and comparison of various retrieval methods [5]. These manually labeled datasets are the basis of search quality analysis and system tuning and against which automated systems are trained and tested. As such datasets are generally constructed through various annotation cycles, quality screening, and adjudication carried out to resolve conflicts, it is a resource-intensive activity that constrains the extent and frequency of data gathering initiatives.

In order to add coverage beyond the abilities of manual annotation, machine learning models are trained using human evaluated data to represent relevance of unlabeled pairs. These models are trained on patterns of text and metadata, and other attributes, allowing groups to estimate human judgment in larger query and product spaces while still being anchored on structured human judgments. The invention of learning-to-rank methods has transformed the way search engine applications utilize labeled data to enhance the quality of ranking. It has been established that learning-to-rank techniques are an important development in information retrieval, which give orderly structures over training ranking models through supervised machine learning methods. These methods can be divided into three primary categories: pointwise methods, which model ranking as a regression or classification task on individual documents; pairwise methods, which learn with relative preferences between document pairs; and listwise methods, which directly optimize ranking metrics using complete result lists. It has been demonstrated that learning-to-rank algorithms can be successfully utilized in order to make the complex patterns of relevance applicable to the complex queries based on their ability to incorporate various features, including query-document similarity scores, document quality indicators, user behavioral cues, and contextual details. Learning-to-rank tutorials and surveys have stressed that, in terms of providing significantly better performance than traditional retrieval models, these techniques have been seen to perform well in areas with rich feature sets and sufficient training data [6]. The trained models allow organizations to make relevance predictions on millions of query-item interactions that could otherwise be impractical to consider manually, and thus compound the value of the original human annotation investment.

The quality and representativeness of the underlying human-labeled training data are, however, fundamental to the effectiveness of these machine learning approaches. Models can only be trained on the patterns that can be found on training sets, and thus, idiosyncrasies, biases, or discrepancies in the human annotations will carry over to model predictions. Also, due to the changing distribution of queries and the changing product catalogs, models trained on the old data can become less precise over time, which requires retraining with new human labels now and then.

4. Limitations of Traditional Methods

However, regardless of their fundamental significance, traditional relevance measurement methods have a major limitation. Human assessment is very costly in terms of annotator training and quality control in order to ensure consistency. Strict procedures are not enough to eradicate inter-rater variability, in which different assessors perceive relevance in diverse ways, which results in label noise. A study of the consistency of human relevance judgments has found that the consensus between annotators may differ significantly with query complexity, ambiguity of results, and the level of the scale of relevance that is being used. Research has reported that despite the clear guidelines and professional assessors, the high inter-annotator agreement is not always possible, especially concerning subtle differences between the levels of relevance. Human judgments can be varied in many ways, and some of the reasons are the variations in domain knowledge, understanding the query intent, individual bias, as well as subjective beliefs regarding what is satisfactorily matched between query and document. The studies of relevance assessment through crowdsourcing have examined the extent to which distributed annotation methods can be used to scale an evaluation process at an affordable cost. Research has demonstrated that although crowdsourcing systems allow reaching wider sources of annotators, it is also accompanied by new difficulties connected to quality regulation, differences in experience among the annotators, and the necessity of efficient aggregation strategies to overcome conflicting decisions. Crowdsourcing the relevance assessment requires proper task design, proper compensation schemes, and quality assurance systems that are capable of isolating and sorting out unreliable annotations [7]. Manual annotation on millions of query-product combinations is effectively impractical, and it gives a coverage gap, especially in the long-tail queries and newly added inventory. The economic and temporal limitations of human annotation imply that organisations have to engage in hard trade-offs between breadth of coverage and depth of annotation, resulting in datasets that can only fully evaluate the most frequent queries and, in the process, leave large portions of query space untested.

Models that are trained with human labels impose the same drawbacks, inconsistencies, biases, and lack of training data directly translate to predictions. Machine learning models can learn to follow systematic bias patterns, or have weak coverage of particular query patterns or product categories, when the training data sets have systematic biases, or do not provide strong knowledge about the principle of relevance. Studies have revealed that the effectiveness of the supervised learning methods in relevance prediction is highly reliant on the quality and variety of the training examples. Research that has been done to assess the performance of the evaluation metrics and techniques has revealed that conventional relevance assessment paradigms should weigh various conflicting goals such as accuracy, consistency, coverage, and practical viability. The difficulty does not just stop at gathering more annotations, but it is also to ensure that the marked data is sufficiently representative of the entire range of query purposes, kinds of products, and relevance connections that occur in actual search settings. Moreover, manually labeled data is sparse, and therefore a large number of combinations of queries and items are never observed in training, so models have to rely on extrapolating limited examples when predicting new queries or products. The evaluation of information retrieval methods has highlighted that the evaluation methods used need to be redesigned to meet the growing scope and complexity of the modern search systems in which traditional methods that rely on exhaustive manual evaluation cannot be viable. Research has shown that better effective evaluation strategies are necessary to give a good quality cue without fully covering all the potential query-document pairs [8]. The high cost, low scalability, and quality issues mandate the development of more efficient, scalable, and adaptable measurement methods that are able to be accurate and measure large evaluation spaces many times over. Organizations need solutions that are capable of producing trustworthy relevance measurements at scale without subsidizing the cost of annotation to prohibitive levels and without compromising the semantic richness required to provide the accuracy of complex query-item relationships.

5. Large Language Models for Relevance Assessment

The relevance judgment capabilities are not advanced in the modern large language models due to their semantic interpretation, contextual, and general knowledge. These models, unlike the traditional methods that were restricted to the keywords match and numeric characteristics, decipher the query intent, comprehend the product descriptions that are not directly obvious, and connect the item relationships in a human-like manner. It has been shown that large language models have amazing capabilities to understand sophisticated linguistic patterns, semantic relations, and contextual subtleties that can allow them to carry out tasks that involve profound knowledge of text. Experiments analyzing the use of such models on the information retrieval problem have demonstrated that the pre-training of such models on large text corpora provides them with an extensive knowledge of the world and reasoning skills that can be used to evaluate the relevance of a query and document. Correspondence of the models in coming up with consistent explanations of their judgments gives them transparency, which traditional black-box ranking models do not have, and this is useful when developers may interpret and debug decisions in relevance assessment. It has been discovered that using large language models appropriately prompted, they can encode fine grained semantic associations, like understanding that a query to get winter clothes may be fulfilled with things characterized by synonymous words or similar ideas, and can be flexible enough to be not strictly tied to single keyword matching.

The process of implementation will start by choosing a suitable relevance framework that will outline the label categories. Explicit mapping policies standardize the line of thought, such as how partial query satisfaction can be a substitute versus irrelevant. Following engineering then takes the model under then processed through systematic evaluation, which uses the chain-of-thought logic, which does not involve making snap judgments, but uses items step by step. Studies that have been done concerning prompting strategies have shown that well-structured prompts have a great impact on model performance as far as complex reasoning is concerned. It has been found that prompting in terms of chain-of-thought, where the models are encouraged to prefigure the inference steps before reaching conclusions, has a significant role in enhancing accuracy on tasks that involve multiple steps of inference. The experiments indicate that, like adding instructions in the form of instructions, such as, let us think step by step, to prompts can significantly improve the performance of the model in arithmetic, common sense, and symbolic reasoning tasks. This method is especially efficient in combination with few-shot learning, where having a limited number of exemplars that are used to illustrate the target reasoning pattern, the model can be generalized to new problems. It has been shown that chain-of-thought prompting helps chain-of-thought models to reach emergent reasoning capabilities in sufficiently large language models, and while the models can solve complex problems, smaller models or models with standard prompting cannot do so effectively [9]. Such a technique is especially useful in relevance evaluation, where the identification of the relationship between a query and an item can be made by taking into account several variables like category correspondence, attribute correspondence, functional likeness, and contextual suitability. The presence of a few-shot examples with the correct labeling patterns is a sign that the model has been shown simple illustrations of the way in which relevancy criteria should be applied in different contexts. Extra information, like product metadata, product specifications, or product reviews, can be used to differentiate between shallowly similar products that contain vital differences that can influence their suitability to particular queries.

Models are tested on high-quality human-labeled datasets before deployment, and accuracy, precision, recall, and systematic patterns of misclassification are measured. The model scales to production only after passing quality checks, accepting query-item pairs in batches, and (optionally) the uncertain cases will be flagged, allowing human inspection. Reliability is further enhanced using techniques of enhancement. Teacher-student architecture refers to large models to produce training data on smaller, faster models that can be deployed in real-time to trade off between reasoning and operational efficiency. Knowledge distillation Research has shown that it is possible to compress a large network of students by training small student networks to emulate the behavior of a large teacher network with significant compression and performance of the original model. It is demonstrated that the student can figure out the dark knowledge within the prediction distributions of the teacher, including subtle relationships not expressible in binary labels, by training the student models on the soft targets generated by the teacher models, without just the hard labels [10]. Human audits periodically identify systematic biases and ensure the maintenance of quality standards.

Table 1: *Relevance Taxonomy Categories and Their Applications in E-Commerce Search [3, 4]*

Relevance Category	Definition	Training Model Signal Strength	Evaluation Metric Clarity	User Intent Alignment
Exact Match	Precisely fulfills the query	Very High	Very High	95%
Substitute	Functionally similar alternative	High	High	75%
Complement	Pairs naturally with the sought item	Moderate	Moderate	60%
Irrelevant	Does not satisfy the query	Low	Very High	5%

Table 2: *Factors Affecting Human Annotation Quality and Inter-Rater Agreement in Relevance Judgment [5, 6]*

Quality Factor	Impact on Agreement Rate	Resource Investment Required	Scalability Limitation	Quality Control Complexity
Annotator Training	High (75-85% agreement)	Very High	Low	Moderate
Guideline Clarity	Very High (80-90% agreement)	Moderate	Moderate	Low
Query Complexity	Negative (-15-25% agreement)	Low	High	High
Taxonomy Granularity	Negative (-10-20% agreement)	Moderate	Moderate	Very High
Multiple Annotation Rounds	High (+10-15% agreement)	Very High	Very Low	High

Table 3: *Coverage-Quality-Cost Trade-offs in Traditional Relevance Measurement Approaches [6, 7]*

Annotation Approach	Query Coverage (%)	Annotation Depth	Cost per Query-Item Pair	Quality Consistency	Scalability to Millions of Pairs
Expert Annotators	15-20	Very High	High ($2-5)	Very High (85-90%)	Very Low
Trained Internal Teams	30-40	High	Moderate ($1-3)	High (75-85%)	Low
Crowdsourcing	50-65	Moderate	Low ($0.10-0.50)	Moderate (60-75%)	Moderate
Hybrid Approach	40-55	High	Moderate ($0.80-2)	High (70-80%)	Moderate

Table 4: *Accuracy Improvements from Chain-of-Thought Prompting in Complex Reasoning Tasks for Relevance Judgment [9, 10]*

Reasoning Task	Standard	Chain-of-Thought	Performance	Few-Shot

Type	Prompting Accuracy (%)	Prompting Accuracy (%)	Improvement (%)	Enhancement (%)
Arithmetic Reasoning	45	78	+33	+12
Commonsense Reasoning	52	82	+30	+15
Symbolic Reasoning	38	71	+33	+18
Multi-Step Inference	41	76	+35	+14
Relevance Assessment	58	85	+27	+10

6. Conclusions

Search relevance measurement in e-commerce settings has long been transformed from less efficient human annotation and noisy behavioral feedback into state-of-the-art automated evaluation by large language models. These models overcome intrinsic limitations of the traditional methods by offering scalable, semantically rich evaluation functionality that is able to process millions of query-item pairs and yet maintain consistency and accuracy. By following organized implementation procedures such as selection of relevance frameworks, explicit mapping rules, chain-of-thought prompting, and strict validation of relevance assessment systems against human-labeled benchmarks, relevant organizations can develop relevance assessment systems that reflect subtle semantic associations and relative suitability that can not be achieved using keyword matching techniques or feature-based strategies. Improvement methods like teacher-student structures allow practical implementation by matching the complex reasoning of large models with the latency needs of production systems, and active learning and periodic auditing are used to guarantee the ongoing quality and equity. This paradigm shift can help search teams eliminate the coverage-quality-cost trade-offs that have traditionally limited the relevance measurement, allowing continuous improvement cycles that can transform search systems into progressively more intelligent and attentive to the various user intents. By leaving the use of implicit cues and partial manual annotations in favor of detailed automated evaluation, organizations can create search experiences that actually take into account user requirements and provide results that meet their intent instead of simply getting them to open their wallets, positioning platform trust and business performance in competitive e-commerce environments more favorably.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Imede Saidi et al., "Entities recommendations using contextual information," ResearchGate, August 2024. Available: https://www.researchgate.net/publication/382776765_Entities_recommendations_using_contextual_information

[2] Peter Bailey et al., "Relevance assessment: are judges exchangeable and does it matter," ResearchGate, July 2008. Available: https://www.researchgate.net/publication/221301256_Relevance_assessment_are_judges_exchangeable_and_does_it_matter

[3] Joachims et al., "Accurately interpreting clickthrough data as implicit feedback," ResearchGate, August 2005. Available:

https://www.researchgate.net/publication/200110530_Accurately_interpreting_clickthrough_data_as_implicit_feedback

[4] Adam Wasilewski, "Harnessing generative AI for personalized E-commerce product descriptions: A framework and practical insights," ScienceDirect, August 2025. Available: https://www.sciencedirect.com/science/article/abs/pii/S0920548925000418

[5] Saba Shaukat et al., "Using TREC for developing a semantic information retrieval benchmark for Urdu," ScienceDirect, May 2022. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457322000619

[6] Hang Li et al., "Learning to rank for information retrieval LR4IR 2009," ResearchGate, December 2009. Available: https://www.researchgate.net/publication/220466684_Learning_to_rank_for_information_retrieval_LR4IR_2009

[7] Omar Alonso et al., "Using crowdsourcing for TREC relevance assessment," ScienceDirect, November 2012. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457312000052

[8] Matteo Palmanori et al., " Aggregated search of data and services," ScienceDirect, April 2011. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306437910000979

[9] Jason Wei et al., "Chain of Thought Prompting Elicits Reasoning in Large Language Models," ResearchGate, January 2022. Available: https://www.researchgate.net/publication/358232899_Chain_of_Thought_Prompting_Elicits_Reasoning_in_Large_Language_Models

[10] Geoffrey Hinton et al., "Distilling the Knowledge in a Neural Network," ResearchGate, March 2015. Available: https://www.researchgate.net/publication/273387909_Distilling_the_Knowledge_in_a_Neural_Network