A Chain-of-Thought Approach to Semantic Query Categorization in e-Commerce Taxonomies

Jetlir Duraj[1], Ishita Khan[1], Kilian Merkelbach[2] and Mehran Elyasi[1]

[1]eBay Inc, USA
[2]eBay Inc, Germany

Abstract

Search in e-Commerce is powered at the core by a structured representation of the inventory, often formulated as a category taxonomy. An important capability in e-Commerce with hierarchical taxonomies is to select a set of relevant leaf categories that are semantically aligned with a given user query. In this scope, we address a fundamental problem of search query categorization in real-world e-Commerce taxonomies. A correct categorization of a query not only provides a way to zoom into the correct inventory space, but opens the door to multiple intent understanding capabilities for a query. A practical and accurate solution to this problem has many applications in e-commerce, including constraining retrieved items and improving the relevance of the search results.

For this task, we explore a novel Chain-of-Thought (CoT) paradigm that combines simple tree-search with LLM semantic scoring. Assessing its classification performance on human-judged query-category pairs, relevance tests, and LLM-based reference methods, we find that the CoT approach performs better than a benchmark that uses embedding-based query category predictions. We show how the CoT approach can detect problems within a hierarchical taxonomy. Finally, we also propose LLM-based approaches for query-categorization of the same spirit, but which scale better at the range of millions of queries.

Keywords

Query Categorization, E-commerce Search, Taxonomies, Chain-of-Thought Reasoning, Large Language Models

1. Introduction

Mapping user queries to relevant categories is essential in e-Commerce search and navigation, since it enhances search relevance, user navigation, and inventory targeting. Traditionally, demand-based methods leveraging user behavioral data like click-through rates have been researched in industry and academic literature. However, these methods face issues like presentation bias and signal sparsity, particularly with long-tail queries and new inventory (e.g., Joachims et al. [1], Xv et al. [2]).

Recent semantic-based methods offer promising solutions by using linguistic and contextual understanding to infer relevance, addressing sparsity and bias while generalizing to new scenarios. These methods integrate query semantics with taxonomies for more precise mappings, but often lack task-specific adaptations and focus on static representations (e.g., Dehghani et al. [3], Gao et al. [4]).

We propose a novel semantic projection system that complements demand-based methods. We adapt the chain-of-thought (CoT) reasoning paradigm for large language models (LLM) (see Wei et al. [5]) to our specific problem of classification in a hierarchical taxonomy. Given a query our system navigates from root to leaf categories of the taxonomy, integrating query semantics and taxonomy details to create precise, interpretable mappings. Our approach is orthogonal to demand-based methods and aims to enrich and de-bias them.

Similar in spirit to chain-of-thought (CoT) reasoning where an LLM solves a complex task by breaking it down into smaller steps and tasks, the system we build operates through a structured, multi-step process towards the solution. It iteratively predicts ranked categories at each taxonomy level, moving from root to leaf categories. Our model dynamically adjusts prediction thresholds based on the semantic information of the current category node, its children, and the query semantics.

jduraj@ebay.com (J. Duraj); ishikhan@ebay.com (I. Khan); kmerkelbach@ebay.com (K. Merkelbach); melyasi@ebay.com (M. Elyasi)

A key feature of our approach is its ability to specify context for the query — such as user intent of buying, browsing, or accessory/complementary intents. Context specification enables more targeted and contextually appropriate category mapping. Our model provides confidence scores for each prediction, thus ranking categories with the aim of offering actionable insights and interpretability. Additionally, it can serve as a diagnostic tool for refining and improving taxonomies, addressing structural noise often misaligned with buyer signals.

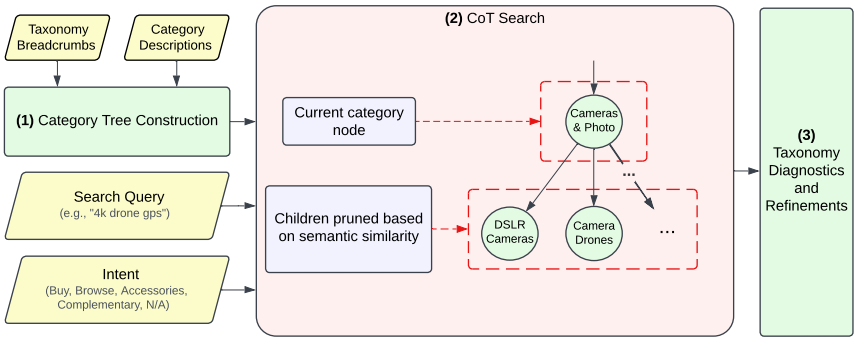

Figure 1: Overview of our system architecture. First, we build our category tree and enrich it with category descriptions. Then, we search within the tree for the node that best fits the search query. Finally, diagnostics and refinements of our taxonomy can be extracted.

1.1. Related Work

Recommendation systems and personalized search in e-Commerce heavily depend on understanding user interaction data semantically. Two main approaches exist in the literature: demand-based methods, using behavioral data, and semantic methods, enhancing recommendations through query and content understanding.

Demand-based approaches to category prediction These approaches use implicit feedback, like click-through data, to personalize recommendations. Ai et al. [6] introduced a model using local contexts for improved ranking, while Joachims et al. [1] highlighted click-through data's utility despite biases. Xv et al. [2] used graph neural networks to better handle long-tail queries and cold-start products, thus tackling the persisting challenge of data sparsity.

Semantic-based methods and query understanding The focus here is on understanding query intent and matching it with relevant content. Guo et al. [7] proposed a deep relevance matching model, and Mitra and Craswell [8] introduced semantic embeddings for aligning queries with documents. These methods help mitigate bias and sparsity issues. Dehghani et al. [3] demonstrated weak supervision's effectiveness in sparse datasets, while Gao et al. [4] explored semantic generalization in taxonomies.

LLMs LLMs are good at generalizing to unseen scenarios, which is a crucial capability for handling long-tail queries in e-Commerce. Brown et al. [9] showed LLMs' strengths in few-shot learning. Chain-of-thought reasoning (Wei et al. [10], Kojima et al. [11]) enhances hierarchical reasoning and predictions.

Taxonomy integration Taxonomies offer a structured basis for improving category projections and recommendations. Huang et al. [12] showed how to integrate semantic signals and taxonomies in e-Commerce search.

2. Methodology

2.1. High level overview

The task of identifying semantically relevant leaf categories is a multi-label classification problem given input data. The labels correspond to the leaf categories of a tree-structured taxonomy. This makes tree path-finding algorithms a natural choice to study. Additionally, assessing the strength of the semantic relationship between a query and leaf categories is crucial for applications. Therefore, we integrate straightforward tree search methods with LLM-scoring for semantic relevance. In LLM-scoring, the LLM is asked to provide a score measuring the strength of the semantic relation between a query and an e-commerce category and its description. LLM-scoring for semantic relevance closely approximates human judgment in our task, as demonstrated in Table 1.

Table 1
Classification performance of LLM scoring (human judgment as ground truth).

LLM	F1	Precision	Recall
Mixtral-8x7B	0.727	0.825	0.649
Llama3-70B	**0.805**	0.797	**0.813**
OpenAI-GPT-4oMini	0.743	**0.862**	0.654

This table presents classification metrics for the semantic relevance of 4,897 query-leaf category pairs, where human judges determine the ground truth. The LLMs we evaluated perform well in terms of F1, precision, and recall. In terms of inference speed, Mixtral-8x7B, a mixture-of-experts model that can be hosted locally, while ranked third in terms of F1, has significantly higher inference speed than the other two models considered.[1] We use Mixtral-8x7B for the results of our methodology in the paper.

Another reason for favoring LLM-scoring for semantic relevance over more generative LLM approaches is the issue of instruction-following. Such issues cannot be entirely eliminated due to the generative nature of current LLM architectures. When prompted to select directly from available children in a category node, LLMs sometimes modify category names rather than reproduce the exact names from inputs. This issue is less prevalent in closed-source LLMs like those from OpenAI or Gemini (4% failure rate in our experiments), but is more pronounced in open-source models from Hugging Face, which can be hosted locally and support large-scale inference.

For these reasons, we focused on a scoring approach for the LLM component of our method: we ask the LLM for a confidence score for semantic relevance, ranging from 1 (lowest) to 10 (highest). The decision to continue searching at each category node is based on the semantic scores of its children, along with other contextual, or algorithm-runtime information.

2.2. Breadth first search design (CoT BFS)

Interpreting the categorization task as one of regression (score assignment) after classification, our approach first solves the classification problem in terms of relevant leaf categories fully, before addressing the regression problem. Specifically, for a given query, at level 1 of the taxonomy, after scoring the semantic relevance of all level-1 categories, we retain only the relatively most semantically relevant children, pruning the rest. We use two query-dependent thresholds for selection: a selection-threshold and a minimum-threshold, both ranging from 1 to 10. The relevance scores (1 to 10) of category children are mapped to the standard normal distribution. The selection-threshold divided by 10 is applied to the standardized scores to prune less relevant children. For example, with a selection-threshold of 9 (out of 10), children scoring below the mean plus 0.9 times the standard deviation of semantic scores are pruned. To deal with potential high skewness of the score distribution at the lower end, a child's

[1]The experiments took less than two hours to run within eBay's infrastructure. OpenAI-GPT-4oMini is closed source, while Llama3-70B is internally hosted.

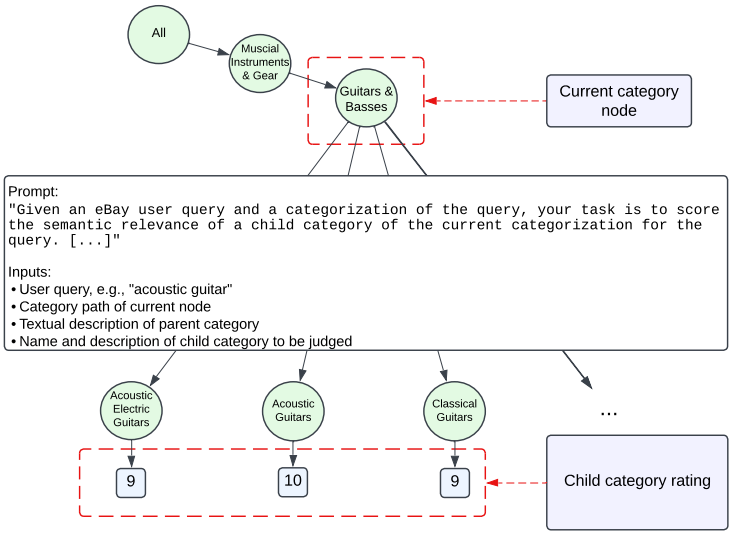

Figure 2: CoT BFS output for the query *acoustic guitar*. Mixtral-8x7B was used as the LLM. The prompt contains information about the path from the root of the tree to the current node, the immediate parent node, and the child node to be rated.

original semantic score must exceed the minimum-threshold (the second threshold) to survive for further exploration.

Next, the algorithm examines each subtree starting from the level-1 children that survived the initial pruning. We repeat the pruning process for children in these subtrees to identify non-pruned second-level nodes using the same relative thresholding procedure described above. This iterative process continues until we reach nodes without children, which are leaf categories, and thus added to the set of candidate leaf categories.

Finally, because the search relies on relative rather than absolute semantic thresholding, we score the final set of leaf categories using only leaf category information (categorization path and descriptions). The surviving leaf categories with high semantic relevance, above the minimum-threshold, are the final predictions.

For our empirical application we choose the selection-threshold and minimum-threshold as follows: given the range of semantic scoring between 1 (lowest) and 10 (highest), we never consider thresholds below 6. Among thresholds 7, 8, 9, for both selection-threshold and minimum-threshold, we only look at pairs where selection-threshold is above the minimum-threshold. Among such pairs, we pick the one that does best in terms of F1-score against a human judgment dataset composed of about 1000 representative queries.[2]

We refer to this method as the Chain-Of-Thought Breadth-First-Search (CoT BFS).

Figure 2 illustrates the CoT BFS categorization result for the query *acoustic guitar*, with selection-threshold of 9, and minimum-threshold of 8. In the first step, CoT BFS narrows down to a single level-1 category: *AllCats* > *Musical Instruments & Gear*, which gets an intermediate semantic score of 10. Out of 35 level-1 categories all other level 1 categories have low score, with the mode score of 1, and maximal other score of 4 (category *AllCats* > *Music*). The surviving level-1 category *AllCats* > *Musical Instruments & Gear* has 16 children. The next classification step prunes out 15 out of these 16 children, because all of them have semantic score lower than 8. The child *Guitars & Basses* has semantic score of 10 and

[2]Note that the maximal score of 10, i.e. the span of the range 1..10 is also a hyperparameter. These hyperparameters need to be validated periodically over time, to account for distribution shifts of the queries, but also changes in the taxonomy. Space constraints preclude us from including a detailed analysis of the effect of the hyperparameters. Here we report qualitatively the following: lowering the selection-threshold and minimum-thresholds typically increases recall, but lowers precision. We also found that using the alternative range 1..5 instead of 1..10 lowered both precision and recall, while using 1..20 resulted in slightly higher recall, lower precision and lower F1.

survives. Further, the node *AllCats > Musical Instruments & Gear > Guitars & Basses* has 13 children. Scoring these children in the next step prunes out all but three children. The surviving children are the nodes: *Classical Guitars* (with a final score of 9), *Acoustic Electric Guitars* (final score of 9), and *Acoustic Guitars* (final score of 10). At this point, the search stops, as the reached nodes are already leaf-categories of the category tree.

2.3. Scalable approaches for CoT and LLM scoring

For a given query, the total number of LLM calls in CoT BFS is in the same order with the number of category nodes visited in the taxonomy. Experiments on large datasets of queries show CoT BFS can visit between 1.7% to 24.8% of the total number of category nodes of eBay's taxonomy. This shows the efficiency of our method, given the very high number of categories in eBay's taxonomy. Nonetheless, to scale this method to millions of queries and low latency, modifications are needed. We propose two approaches, the second more scalable than the first.

2.3.1. CoT-k-NN hybrid BFS

k-NN retrieval based on embeddings of category names or descriptions can be used as a filter at each step of the tree search process. Instead of exhaustively rating each child node, only a subset surviving the embedding distance filter (between the user query and a textual representation of the category) is scored by the LLM. This reduces the number of LLM calls at each node of the taxonomy, and constrains the search to only the most promising directions.

2.3.2. k-NN-search + LLM scoring on leaf categories

One replaces tree-search with a k-NN-search on leaf category embeddings as a pre-filter, followed by LLM scoring of the candidates identified through k-NN. Running k-NN with many neighbors at the beginning of the procedure, e.g. 20 neighbors, enhances recall. We use a variant of this method in section 3.2 to construct a synthetic ground truth for evaluating CoT BFS.

3. Experimentation

3.1. Baseline model: k-nearest neighbors categorization

Our benchmark for evaluation of the CoT BFS is k-NN search for leaf categories with $k = 10$ using (not fine-tuned) embeddings from sentence-BERT. Cosine-similarity is used as a metric for the k-NN.[3] We provide detailed categorization performance, comparing our method's F1, precision and recall classification metrics in the micro, macro and sample aggregations. Micro aggregation considers performance across all queries and leaf categories. Sample aggregation considers performance per-query and then aggregates. Macro aggregation considers performance per leaf category and then aggregates.

3.2. Evaluation against baseline

3.2.1. Human Judgment

Human judgment offers both qualitative and quantitative evaluations by utilizing human intuition and expertise. Evaluators review predicted categories for semantic relevance, though this process is subjective and costly for large datasets. Despite its costs, human judgment captures nuances often missed by other methods, hence it is indispensable. Our human judgment dataset includes 1018 queries

[3]We have access to language models trained on eBay-specific data, that typically perform better in eBay-related tasks than general-purpose language models, but we do not present results from the use of eBay-specific language models. This is because our focus is on understanding how the CoT BFS approach performs with general-purpose, non-fine-tuned LLMs. Furthermore, using language models that are publicly accessible helps with the reproducibility of the results.

and 4897 query-category pairs, judged on semantic relevance (Yes/No decision). We note that the leaf categories for judgment were chosen based on user behavior signals, leading to presentation bias influenced by eBay's current models in production. The annotators are three eBay-funded independent domain experts for eBay's taxonomy.

Table 2 shows the relative performance of the CoT BFS to the benchmark, assuming that the ground truth is given by the human judgment. CoT BFS outperforms the baseline model, especially in relation to the F1 score and precision.

Table 2
Classification Results on Human judgment data. CoT BFS Select-threshold 9, min-threshold 8. Relative improvement to benchmark.

Metric	F1	Precision	Recall
Micro Aggregation	+89.8%	+86.1%	+3.4%
Sample aggregation	+109.7%	+190.6%	-13.5%
Macro aggregation	+44.7%	+60%	+31.4%

3.2.2. AI Pseudo-Reference Method

We use an AI pseudo-reference method to create a dataset that approximates ground truth without the presentation bias found in demand signal datasets. Starting with 3000 user queries, a high-quality LLM emulates human judgment on semantic relevance for query-category pairs. To avoid losing potentially relevant leaf categories for LLM-scoring, we use a k-NN embedding-based search with a large number of neighbors. Namely, we pick out the 100 most relevant categories for each query. From these, we exclude those with cosine similarity below 0.01.

Afterwards, each pair is scored from 1 to 10 using a superior LLM (OpenAI-GPT-4o-Mini) compared to locally hosted Mixtral-8x7B we use for CoT BFS. This hybrid method with large k, see also subsection 2.3.2, delivers a proxy for ground truth. Table 3 depicts the results. CoT BFS again outperforms the baseline, especially in terms of F1 and precision.

Table 3
Classification Results on LLM-judged Data (OpenAI-GPT-4oMini). CoT BFS Select-threshold 9, min-threshold 8. Relative improvement to benchmark.

Metric	F1	Precision	Recall
Micro Aggregation	+96.7%	+137.5%	-5.4%
Sample aggregation	+132.1%	+89.8%	+3.7%
Macro aggregation	+112.1%	+47.8%	+111.4%

3.2.3. Retrieval Test

The retrieval test evaluates predicted leaf categories by comparing recall and relevance between our model and the baseline at the level of retrieved items. Items from the inventory are retrieved based on leaf categories, with estimated recall size showing the proportion of relevant items found. Relevance is measured using an eBay-internal PEGFB model that has been trained on human judgment data, and which classifies results into five graded relevance levels: Perfect, Excellent, Good, Fair, and Bad. The retrieval test evaluation highlights the model's practical utility in improving user satisfaction and search efficiency. Our model significantly outperforms the k-NN benchmark in both recall and relevance, with Mann-Whitney U test results showing highly significant differences in favor of CoT BFS.

Table 4
Classification Results for 3000 queries. CoT BFS Select-threshold 9, min-threshold 8. Relative improvement to benchmark.

	Estimated Recall Size	Relevance Score
Mean	+72%	+34%
Mann-Whitney U stat	5401539.0	5363320.0
Mann-Whitney U p-value	8.95e-88	1.30e-82

4. Applications

4.1. Context learning

By learning from extensive real-world datasets, LLMs can identify patterns that reveal user intent and preferences, enabling personalized search leading to higher semantic relevance. Contextual learning can refine model outputs based on a given context of the query, ensuring that prediction results are relevant. This capability is important for platforms like eBay, where discerning buyer intent enhances the search experience. We consider two applications of context learning, one on user intent and one on brand origin.

More specifically, eBay's taxonomy includes accessory-related categories across various merchandise segments like electronics, automotive, and fashion. The CoT BFS approach can easily incorporate buyer intent by modifying LLM prompts to include intents such as *buying*, *seeking accessories*, *looking for complementary items*. Figure 3 illustrates for the query *canon camera*.

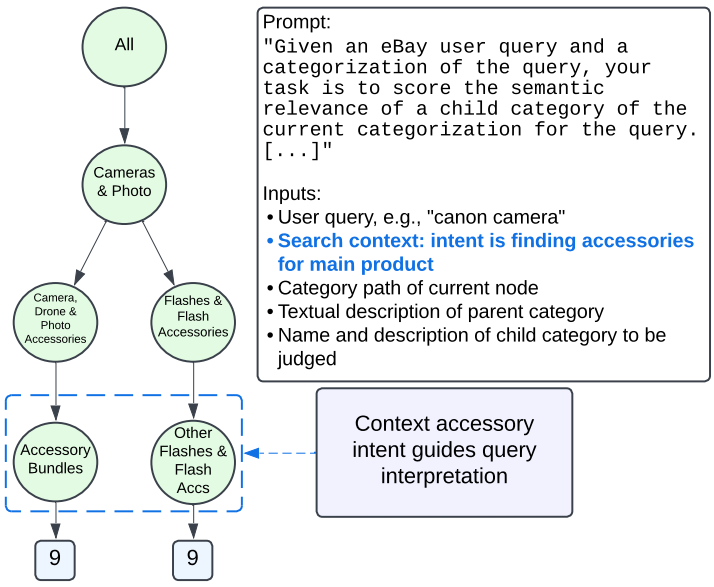

Figure 3: CoT BFS output for the query *canon camera* with accessory intent. The query intent being supplied as context guides the category search into categories that correspond to accessories for the entity in the query.

Without intent, the top category identified for the query is *AllCats > Cameras & Photo > Digital Cameras* with a score of 10. By injecting *accessory* intent as a search context into the prompt, the top categories identified are *AllCats > Cameras & Photo > Camera, Drone & Photo Accessories > Accessory Bundles* and *AllCats > Cameras & Photo > Flashes & Flash Accessories > Other Flashes & Flash Accs*, with a score of 9 each. More generally, table 5 shows how the average semantic scores for Accessory vs. No-Accessory predicted categories change for 15 selected queries, when specifying *accessory* intent.

Including buyer intent in CoT BFS leads to better targeting of relevant categories.

Similarly, we consider the effect of injecting context regarding brand origin. To illustrate, for the

Table 5
Context learning: Accessory Intent Improves Prediction for Accessory Categories.

Acc. Category? (Y/N)	Avg. Score for No Intent	Avg. Score for Acc. Intent
No	**9.143**	1.000
Yes	2.643	**7.214**

query *sports car*, we consider the two distinct contexts of *brand origin is from Germany*, and *brand origin is from Italy*. For this query, the CoT BFS predicts exclusively sports car-related leaf categories from the german brands Audi, BMW, Mercedes-Benz, Porsche in the first case, and exclusively leaf categories from the italian brands Alfa Romeo, De Tomaso, Ferrari, Fiat, Maserati, Lamborghini in the second.

4.2. Detecting issues and improving the taxonomy

By analyzing query patterns and identifying gaps in category representation, CoT BFS can help provide actionable insights for improving e-Commerce taxonomy structures.

In this regard, we conducted an experiment using a representative sample of 25,000 queries processed through the CoT BFS approach with high thresholds: selection-threshold of 10, minimum-threshold of 9. There were 3,110 queries with empty model predictions, indicating that the current eBay taxonomy lacks category nodes at the first few levels that are strongly semantically related to these queries. We uncovered certain patterns by clustering these queries using k-NN search on embeddings. For instance, two clusters of these "failing" queries correspond to the e-Commerce categories *Designer Sunglasses* and *Optical Instruments and Accessories*. The closest leaf categories in the current eBay taxonomy for the first cluster identified (*Designer Sunglasses*) are *AllCats > Clothing, Shoes & Accessories > Women > Women's Accessories > Sunglasses & Sunglasses Accessories > Sunglasses* and *AllCats > Clothing, Shoes & Accessories > Men > Men's Accessories > Sunglasses & Sunglasses Accessories > Sunglasses*, both in a depth of 5 in the taxonomy. A similar issue is observed for the second cluster. These types of insights, when drawn from large sets of user queries, can help product management teams in their taxonomy enhancements work. E.g., introducing a level-2 category titled *Optical Products and Eyewear* with subcategories such as *Designer Sunglasses* and *Optical Instruments and Accessories* might be beneficial for the search experience.

Ultimately, maintaining an e-Commerce taxonomy that provides high value to users involves complex business and product management decisions. Our methodology offers tools to explore taxonomy issues with the goal of enhancing decision making in these complex business decisions.

5. Conclusion and future work

In this study, we introduce a novel methodology for query categorization within hierarchical taxonomies. It combines the world knowledge of LLMs and simple tree search algorithms to achieve high-quality categorization and provide deep insights into the taxonomy.

AB-tests are planned for the scalable methods presented in section 2.3. These involve direct tests where the model predictions are cached for use in production, but also indirect tests where the AB test is on lower-latency categorization models trained with data that have been LLM-labeled via CoT BFS.

Further, we developed a CoT algorithm version that uses absolute thresholding at each taxonomy node, rather than the relative thresholding discussed in the paper, here left out due to space constraints. This method, called Chain-of-Thought Depth-first-search (CoT DFS), searches for leaf categories in a depth-first manner and halts a path when encountering an intermediate node with low absolute semantic relevance, as opposed to relative described in this paper. Because of its more stringent requirements, the CoT DFS approach leads to more queries with empty predictions. CoT DFS can leverage user query activity and LLM semantic-knowledge more effectively than CoT BFS for the purpose of taxonomy diagnostics.

References

[1] T. Joachims, L. Granka, B. Pan, H. Hembrooke, G. Gay, Accurately interpreting clickthrough data as implicit feedback, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 154–161. URL: https://doi.org/10.1145/1076034.1076063. doi:10.1145/1076034.1076063.

[2] G. Xv, C. Lin, W. Guan, J. Gou, X. Li, H. Deng, J. Xu, B. Zheng, E-commerce search via content collaborative graph neural network, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2885–2897. URL: https://doi.org/10.1145/3580305.3599320. doi:10.1145/3580305.3599320.

[3] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, W. B. Croft, Neural ranking models with weak supervision, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 65–74. URL: https://doi.org/10.1145/3077136.3080832. doi:10.1145/3077136.3080832.

[4] J. Gao, P. Pantel, M. Gamon, X. He, L. Deng, Modeling interestingness with deep neural networks, in: Conference on Empirical Methods in Natural Language Processing, 2014. URL: https://api.semanticscholar.org/CorpusID:2141094.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 24824–24837.

[6] Q. Ai, K. Bi, J. Guo, W. B. Croft, Learning a deep listwise context model for ranking refinement, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 135–144. URL: https://doi.org/10.1145/3209978.3209985. doi:10.1145/3209978.3209985.

[7] J. Guo, Y. Fan, Q. Ai, W. B. Croft, A deep relevance matching model for ad-hoc retrieval, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 55–64. URL: https://doi.org/10.1145/2983323.2983769. doi:10.1145/2983323.2983769.

[8] B. Mitra, N. Craswell, An introduction to neural information retrieval, Foundations and Trends® in Information Retrieval 13 (2018) 1–126. URL: http://dx.doi.org/10.1561/1500000061. doi:10.1561/1500000061.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[11] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, ArXiv abs/2205.11916 (2022). URL: https://api.semanticscholar.org/CorpusID:249017743.

[12] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck, Learning deep structured semantic models for web search using clickthrough data, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 2333–2338. URL: https://doi.org/10.1145/2505515.2505665. doi:10.1145/2505515.2505665.