DIVAgent: A Diversified Search Agent that Mimics the Human Search Process

Zhirui Deng
zrdeng@ruc.edu.cn
Renmin University of China
Beijing, China

Jingfen Qiao
j.qiao@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Zhicheng Dou
dou@ruc.edu.cn
Renmin University of China
Beijing, China

Ji-Rong Wen
jrwen@ruc.edu.cn
Renmin University of China
Beijing, China

Maarten de Rijke
m.derijke@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Abstract

Search result diversification plays a crucial role in addressing query ambiguity and multi-faceted information needs by reducing redundancy across documents. While previous supervised approaches can achieve superior performance, they require costly, large-scale annotated data. In contrast, unsupervised methods are more flexible and training-free but rely on manually designed ranking functions, often leading to suboptimal performance. Inspired by how humans explore diverse information during real-world searching, we propose a diversified search agent `DIVAgent` to combine the advantages of supervised and unsupervised methods. `DIVAgent` introduces LLMs as the "brain" to reason over complex and diverse search results and delineate human cognitive processes into a workflow tailored for search result diversification. Our search agent first identifies potential user intents and then analyzes the alignment of each document to the intents via an intent-aware module. To guide the generation of diversified document rankings, we design an intent-guided ranker that explicitly links documents to their dominant intents while performing greedy document selection. Experimental results demonstrate that `DIVAgent` significantly outperforms existing unsupervised baselines and achieves competitive performance with supervised models, highlighting the promise of LLMs for diversified ranking in realistic search scenarios.

CCS Concepts

• **Information systems** → **Information retrieval diversity**.

Keywords

Large language model; Search agent; Diversification

ACM Reference Format:
Zhirui Deng, Jingfen Qiao, Zhicheng Dou, Ji-Rong Wen, and Maarten de Rijke. 2025. DIVAgent: A Diversified Search Agent that Mimics the Human Search Process. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761059

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761059

1 Introduction

Information diversification plays a vital role in the information-seeking process of humans as it addresses two long-standing challenges: the inherent ambiguity of short queries and the heterogeneous information needs of different users when issuing identical queries. Traditional search engines typically focus on relevance alone and deliver homogeneous search results for a specific query, overlooking the diverse information needs of users [12, 13, 16, 18, 34, 36, 38]. To bridge this gap, search result diversification approaches have been devised to present diverse documents covering various subtopics, and have become indispensable in both conventional search engines [20, 32, 35, 39, 50, 54] and emerging retrieval-augmented generation (RAG) systems [15].

Initial efforts for search result diversification focus on **unsupervised approaches** [4, 9, 35] that use a hyperparameter, λ, to balance relevance to the query and distinctness between documents, visualized in Figure 1(a). These methods, while straightforward and without need for training, heavily rely on manually defined functions with empirically tuned hyperparameters and typically yield weaker diversity gains. Consequently, research has shifted towards **supervised learning** [14, 42, 47, 50, 53, 54]. As shown in Figure 1(b), these methods construct approximate ideal rankings as ground-truth rankings and automatically optimize diverse ranking functions, leading to superior diversified ranking quality. Nevertheless, supervised learning approaches demand large-scale and high-quality labeled training data. Constructing such data is costly and labor-intensive. E.g., in the widely used ClueWeb09 dataset, human effort is needed to mine potential user intents for each query from query logs and to annotate document relevance corresponding to each user intent. This leads to an important research question: *Can we combine the advantages of unsupervised and supervised methods to develop a method that is simple and with minimal human involvement, while ensuring effectiveness and transferability?*

Large language models (LLMs) [19, 29, 44] offer a promising solution to this problem. With strong zero-shot and few-shot generalization capabilities, they have attracted growing interest in using their zero-shot language understanding and reasoning capabilities in the information retrieval (IR) domain [26, 52]. Most

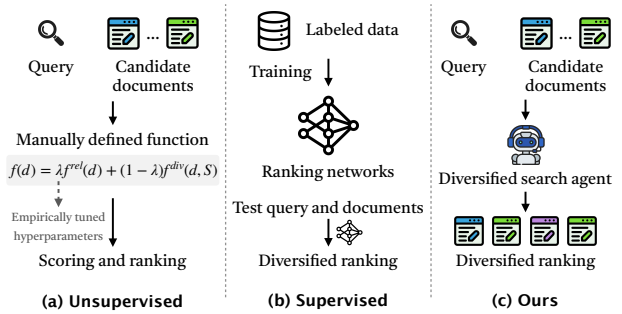

Figure 1: The comparison between our diversified search agent and previous unsupervised and supervised approaches.

approaches [43] focus on exploiting LLMs for relevance ranking, neglecting the diversity of ranking results. Compared to conventional relevance ranking, search result diversification involves striking a delicate balance between relevance and the coverage of diverse user intents. This dual objective introduces additional complexity to the ranking process. Further complicating this problem is the fact that LLMs are not explicitly trained to optimize for diversity during pretraining. Consequently, LLMs inherently lack a proper understanding of diversity in the context of IR. Therefore, effectively using the LLMs' capabilities for diversified document ranking remains a challenging and unexplored problem.

Considering the aforementioned concerns, we take inspiration from human search behavior [2, 21, 25, 37]. When people confront complex and diverse information from a search engine, they can effectively analyze and extract information to satisfy their diverse information needs. E.g., when a user issues a query such as *London travel guide*, she may simultaneously seek information about *tourist attractions*, *local cuisine*, and *public transportation*. To fulfill these diverse needs, users typically engage in a diversified exploration process: they begin by internally formulating specific information searching goals based on their latent intents, then sequentially scan the search engine results, selectively engaging with documents that introduce novel and intent-relevant aspects. Redundant or overlapping content is often skipped. This greedy selection process and the intrinsic intent comprehension capability address the previous challenges posed by integrating LLM in search result diversification.

We propose a **Div**ersified search **Agent** (`DIVAgent`) that mimics human search behavior to produce diversified document rankings. We use LLMs [1, 19, 29, 44] as the "brain" of our search agent, because of their strong comprehension and superior performance in various scenarios without task-specific fine-tuning. To tailor LLMs for search result diversification, we design an agent architecture with three key modules: an intent-aware module, a memory module, and an intent-guided ranker module. When receiving a query, the **intent-aware module** identifies potential user intents underlying the query. Then, the agent iteratively examines each candidate document to determine which intents it satisfies. Key information from these assessments is stored in the **memory module** for subsequent processing. To guide the search agent to produce a more diversified ranking, we devise an **intent-guided ranker module**. This module generates the final document ranking while explicitly associating each document with its dominant intent, thereby enhancing the diversity and interpretability of the search results.

Experimental results demonstrate that `DIVAgent` can significantly outperform existing methods without task-specific fine-tuning and also achieve competitive results compared with supervised approaches. This indicates the benefit of mimicking human search behavior for building a diversified search agent. We further conduct comprehensive analyses to investigate the efficiency and performance of `DIVAgent` with different experiment settings, validating the robustness and wide applicability of our search agent.

Our main contributions are three-fold:

(1) We propose a diversified search agent, `DIVAgent`, that uses LLMs to enhance the diversity of document ranking. This approach combines the strengths of unsupervised and supervised methods, achieving effective performance without relying on labeled data and task-specific fine-tuning.
(2) We take inspiration from human search behavior and devise a three-stage workflow, offering a systematic and explainable way to identify potential user intents and assess the relevance between documents and these intents.
(3) We introduce an intent-guided ranker module that not only generates the final document ranking but also explicitly associates each document with its covered intents. This design enhances the search agent's ability to produce more diversified and interpretable rankings.

2 Related Work

2.1 Search Result Diversification

Search result diversification approaches can be divided into unsupervised and supervised approaches.

Unsupervised Approaches. Pioneering unsupervised search result diversification approaches date back to Maximum Marginal Relevance (`MMR`) [4], which introduced a hyperparameter λ to balance the query-document relevance and diversity of documents. Following `MMR`, `xQuAD` [35] used sub-queries representing pseudo user intents and diversified document rankings by directly estimating the relevance of the retrieved documents to each sub-queries. `PM2` [9] further optimized proportionality by iteratively determining the topic that best maintained the overall proportionality. Subsequently, `HxQuAD/HPM2` [17] extended the concepts of `xQuAD` and `PM2` by incorporating hierarchical subtopics to better model user intents, while `TxQuAD/TPM2` [10] focused on directly modeling term-level subtopics, addressing the challenge of subtopic mining.

Supervised Approaches. Supervised approaches automatically learn the ranking functions for search result diversification by building approximate ideal rankings as ground-truth rankings. `R-LTR` [54] treated search result diversification as a learning-to-ranking problem and optimized ranking functions with constructed ground-truth rankings. To directly optimize evaluation metrics, `PAMM` [47] devised a maximal marginal relevance model for ranking, while `DALETOR` [50] proposed diversification-aware losses to approach the optimal ranking. `NTN` [48] further automatically learned a nonlinear novelty function for measuring the subtopic coverage of documents. With the advancement of deep neural networks, `DSSA` [20] proposed a list-pairwise loss for effectively diversifying document ranking. Moreover, `DESA` [32] employed the attention mechanism to model the novelty of documents and the explicit subtopics. Based on `DESA` [32], `GDESA` [33] incorporated greedy

document selection to approach global optimal ranking results. To address the issue of high-quality training samples shortage, DVGAN [22] adopted Generative Adversarial Networks (GANs) to produce more training samples efficiently while CL4DIV [14] integrated contrastive learning for learning better initial document representation. Graph4DIV [39] and KEDIV [40] further used graph neural networks for measuring the intent coverage differences among documents. To model subtle document subtopic coverage, HAD [11] proposed a hierarchical attention framework to combine intra- and inter-document interactions while PAD [41] segmented the entire document into multiple passages for passage-aware interaction. Besides, DUB [51] introduced an aspect extractor to enhance the intrinsic interpretability and effectiveness of the model.

Previous unsupervised approaches are straightforward but demonstrate inferior diversity improvements. In contrast, supervised approaches perform better but are constrained by the scarcity of labeled data. In this paper, we take inspiration from the human information-seeking process and devise a search agent to directly diversify document ranking without task-specific fine-tuning.

2.2 LLM Search Agents

The abilities of large language models (LLMs) have attracted the interest of researchers to explore LLMs in information retrieval. Pioneering work dates back to WebGPT [28], which adopted LLMs to automatically interact with search engines for answering open-ended questions. MindSearch [6] further introduced a multi-agent framework to solve information-seeking and integration tasks in a web scenario. Another type of work [24, 43] directly explores LLMs for list-wise ranking. RankGPT [43] proposed a prompt-based framework that uses ChatGPT for zero-shot relevance passage ranking. RankVicuna [30] and RankZephyr [31] distill the ranking capabilities of ChatGPT or GPT-4 into moderate-size LLMs.

Different from these methods that solely focus on query-document relevance, in this work, we devise a diversified search agent, aiming to better satisfy the diverse information needs of users.

3 Methodology

Search result diversification has emerged as an effective approach to enhance user satisfaction by providing information covering various aspects. While previous supervised methods demonstrate superior performance, they are heavily constrained by the scarcity of high-quality labeled data. Conversely, unsupervised methods typically depend on manually devised functions, limiting their adaptability and generalizability. Inspired by how humans intuitively select diverse content during real-world information seeking, we propose a diversified search agent DIVAgent to combine the advantages of both paradigms. DIVAgent introduces LLMs as the "brain" to reason over complex and diverse information and delineate human search processes into three modules specialized for diversification. These modules work together to accurately identify "user intents" and fine-grained "document intent coverage" and to effectively diversify document rankings. The architecture of our search agent is depicted in Figure 2.

3.1 Problem Formulation

To begin, the task of search result diversification can be defined as follows. Given a current query q and its initial ad-hoc ranked list $\mathcal{D} = \{d_1, \ldots, d_n\}$ that contains n candidate documents, search result diversification models re-rank these documents and generate a diversified document ranking list $\mathcal{R}$, in which novel documents are ranked higher and redundant ones are ranked lower.

Since enumerating all possible diversified document lists is an NP-hard problem, typically, previous methods either iteratively select the most novel and relevant document or simultaneously score all documents. In this paper, we incorporate the two strategies and devise an LLM-driven diversified search agent.

3.2 Intent-Aware Module

The primary challenges of search result diversification lie in two aspects: (i) accurately identifying diverse user intents underlying search queries, and (ii) effectively recognizing the intent coverage of documents. While LLMs have demonstrated proficiency in relevance ranking tasks [24, 43, 52], directly achieving diversified ranking poses significant challenges. The pretraining objective of LLMs—next token prediction—biases them toward dominant linguistic patterns and statistical co-occurrences, potentially overlooking less frequent user intents and influencing the assessment of document intent coverage.

Observations of human search behaviors reveal that individuals naturally employ implicit reasoning and contextual understanding when navigating non-diversified result lists [21]. They typically strategically select previously unseen and novel content to collectively satisfy their diverse latent information needs. Drawing inspiration from this observation, we mirror the human behavior and introduce an intent-aware module to swiftly detect potential user intents and effectively model document intent coverage, facilitating the subsequent diversifying document ranking process.

3.2.1 User Intent Identification. Recognizing potential user intents is a crucial step for search result diversification. When a query encompasses multiple user intents, search engines should provide diverse results addressing each user intent. In other words, breaking down queries into multiple user intents enables our search agent to satisfy different information needs and increase the comprehensiveness of document rankings. Previous research has primarily relied on commercial search engine query logs [45] or query suggestions [20, 32] to construct user intents, which presents scalability challenges. Noticing the remarkable generalization capability of LLMs, we devise a user intent identification component that allows our search agent to automatically identify the potential user intents associated with a given query.

To ensure the completeness of the decomposed user intents, our search agent is prompted with the following instruction with several demonstrations tailored to extract distinct potential user intents concerning the current query q:

Prompt 1. *Analyze the given query and documents and identify up to 10 distinct user intents when users issue the query. Each user intent should be independent, self-contained, and highlight a unique aspect of the query. Ensuring no semantic overlap among user intents.*

Following the above instruction, the LLM will generate a user intent list $\mathcal{I} = \{i_1, \ldots, i_p\}$ for each query, which will be reused in the subsequent ranking stage.

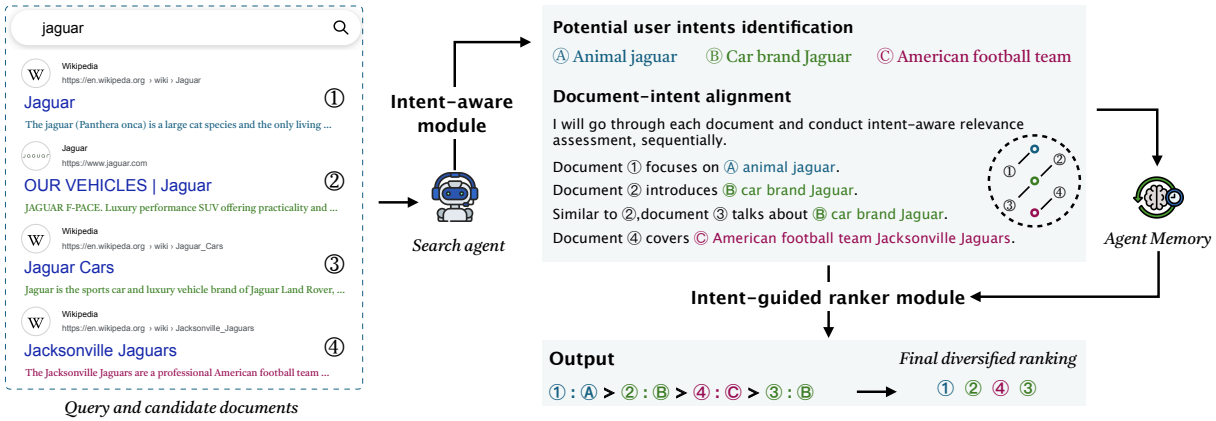

Figure 2: Overview of DIVAgent. Our search agent initially engages with the intent-aware module to identify user intents and reason about document intent coverage. Task-related information is stored in working memory and passed to the intent-guided ranker module for greedy-based document selection and intent-guided output.

3.2.2 Document-Intent Alignment. Given the identified potential user intents, our search agent requires effectively analyzing the intent coverage of each candidate document, so as to support subsequent diversified document ranking. A straightforward approach is to pass each intent with each document to our LLM-based search agent for judging their relevance. But this method leads to substantial computational cost due to repeated model invocations. More importantly, it fails to capture relative content differences across documents, which is critical for fine-grained document-intent alignment. We adopt a more efficient strategy: feeding all candidate documents into the search agent simultaneously and prompting the model to assess their intent coverage comparatively and sequentially. This joint processing allows the model to better understand the intent coverage and semantic distinctions among candidate documents, resulting in a more comprehensive alignment. Nevertheless, processing the full content of all documents can be resource-intensive and may even exceed the context length limits of LLMs. Even long-context LLMs often struggle to handle such extensive input effectively. To tackle this problem, we first propose two strategies that empower the search agent to grasp the key concept of each document while ensuring an efficient document-intent alignment.

Direct Content Extraction. We use the initial segment of each document—typically the first N tokens—as its representative content. This strategy is inspired by the widely adopted principle in web design [27], where critical information is intentionally front-loaded to attract user attention and improve its discoverability. Prioritizing the beginning of a document allows our search agent to efficiently assess intent coverage while adhering to computational and context-length constraints.

Core content Compression. While the beginning of a document often contains high-density information, long documents may address multiple user intents that appear beyond the initial segment. To capture such content while mitigating context-length limitations, we design a core content compression mechanism that generates concise, intent-aware summaries for each document. We employ our search agent to distill the intent-relevant aspects of the document via a prompt-based method, allowing for thorough document comprehension without exceeding computational limits.

After extracting the core content from each document, we instruct our search agent to go through each candidate document and evaluate its intent coverage. To enhance this process, we encourage the agent to simulate a human-like deliberation process through explicit reasoning. Such explicit reasoning facilitates reducing hallucination, improving factual consistency, and leading to more faithful and interpretable document-intent alignment. The prompt used to elicit this reasoning behavior is illustrated below:

Prompt 2. *Assign an intent_id to each document based on the intent it most closely addresses (starting from 1). If a document does not relate to any user intent, assign intent_id to 0.*

3.3 Memory Module

Drawing inspiration from the memory structure of the human brain, we integrate both long-term memory and working memory into the memory module for effectively storing and managing information.

3.3.1 Long-term Memory. For people, long-term memory stores accumulated knowledge, enabling proper reasoning and decision-making based on prior experience. Analogously, in our diversified search agent, long-term memory is embodied in the pre-trained knowledge encoded within the LLMs. This memory encompasses world knowledge and linguistic patterns that play a foundational role in enabling the agent's zero-shot and few-shot generalization capabilities. Our search agent uses such knowledge to identify user intent and analyze document intent coverage, without requiring task-specific supervision.

3.3.2 Working Memory. The working memory serves as a crucial cognitive component in the human brain, responsible for the storage of information related to ongoing tasks. For DIVAgent, the working memory typically collaborates with the subsequent ranker to perform the final diversified document ranking. To guarantee that the ranking results are comprehensive and aligned with the diverse information needs of different users, the working memory needs to provide the following two crucial dimensions of information: (i) The current query q and the initial document ranking $\mathcal{D}$ constitute the environment information and serve as the fundamental basis for the whole diversification process. (ii) For each query,

user intents $\mathcal{I}$, along with the relevance of each candidate document to identified user intents (cf. Section 3.2), will be temporarily saved in working memory as sensory information for executing the following diversified ranking process.

3.4 Intent-Guided Ranker Module

The ranker serves as the central scoring component, determining the order of candidate documents. For search result diversification, the primary objective is to ensure the top-ranked documents capture a broader range of user intents. In this section, we devise an intent-guided ranker module for balancing relevance and diversity and ultimately achieving optimal diversified document rankings.

3.4.1 Greedy-based Selection. As illustrated in Section 3.1, search result diversification confronts the NP-hard challenge, complicating the optimization of diversified results. Greedy-based selection strategies provide a possible solution to this issue [20, 39]. Additionally, LLMs often encounter the lost-in-the-middle phenomenon [23] when processing lengthy content in a single turn, adversely impacting performance. To mitigate these challenges, we propose a greedy-based selection mechanism to iteratively select the next document considering both novelty and relevance, thereby achieving the optimal diversified document ranking.

Initially, given a query q, a query-related candidate document list $\mathcal{D} = \{d_1, \ldots, d_n\}$, and the sensory information (*i.e.*, user intents $\mathcal{I}$ and document-intent relevance), we prompt the search agent to select the first document $d_1^s \in \mathcal{S}$ based on the relevance to the query. For the following document selection, it is essential to consider both the document's relevance to the query and its distinction from already selected documents. Our search agent will repeat the above greedy-based selection steps until selecting the top-K documents $\mathcal{S} = \{d_1^s, \ldots, d_K^s\}$ for each query. The whole process is conducted with the following instructions:

Prompt 3. *Re-rank the document based on two criteria: the diversity of user intent and relevance to the query. 1. Start by selecting the document most relevant to the query. 2. For each subsequent document, select one that is relevant to the query and introduce new user intents not covered by previously selected documents. If no new intents are available, select documents purely by relevance to the query. 3. Ensure that the final list is reorganized to reflect this selection process.*

3.4.2 Intent-Guided Output. To explicitly promote the diversity of document ranking, we devise an intent-guided output format, where each document is tagged with the predicted user intent identifier, *i.e.*, *doc_id:intent_id*. The intent labels act as soft constraints during reranking, encouraging novelty and penalizing documents aligned to the same intent. This leads to rankings with less content overlap and more novel information per position. Moreover, the structured output format provides interpretable results, enhancing explainability, transparency, and ultimately facilitating user trust in the system. We follow [43] and use the symbol > to guide the search agent to directly generate the diversified document ranking order without producing an intermediate score. For example, $(1 : A) > (2 : B) > \cdots$, where the number means the document identifier and the letter indicates the intent identifier.

3.5 Specialization Distillation

Although closed-sourced LLMs demonstrate remarkable capabilities, their deployment is expensive and impacted by high latency. To avoid these problems, we propose specialization distillation, which aims to distill the diversified ranking capability of powerful LLMs into smaller, deployable models.

Formally, given a query q and n candidate documents, we instruct `DIVAgent`, powered by more capable LLMs (*e.g.*, Claude 3.7 Sonnet), to generate the diversified ranking results based on the previously introduced reasoning process. The ranking results are formatted as Section 3.4.2, *i.e.*, $(r_1 : i_1) > \cdots > (r_n : i_n)$, where r_k indicates the document ranked at position k and i_k means the intent covered by r_k. This output is used as the supervision signal for the specialization distillation. Considering the limited learning capacity of smaller models, directly distilling the reasoning process to them is challenging and may lead to capability collapse. Therefore, the objective of the student model $f_\theta(q, \mathcal{S}, d^+, d^-)$ is to determine which document should be prioritized given the selected document set $\mathcal{S}$. (d^+, d^-) is a document pair such that the teacher ranks d^+ higher than d^-. The ranking score is then defined as the generation probability of the document identifier. We train the student model using a list-pairwise loss [20] based on the relative ranking provided by the teacher. The definition of the loss is as follows:

$$\mathcal{L}_{\text{list-pairwise}} = \sum_{q \in Q} \sum_{o \in O_q} |\Delta M|(y_o \log(Z) + (1 - y_o) \log(1 - Z)), \tag{1}$$

where o is the sample pair in all sample pairs O_q of query q, $Z = 1/(1 + e^{-(s_d^+ - s_d^-)})$, s_{d^+} and s_{d^-} are the ranking score of document d^+ and d^-, respectively, and M reflects the margin in utility metrics (*e.g.*, α-nDCG) between the pair.

4 Experiments

4.1 Datasets and Evaluation Metrics

In light of the limited datasets suitable for search result diversification, we use the ClueWeb09 Category B data collection [3] for our experiments, aligning with prior research [20, 32, 39].[1] The ClueWeb09 dataset comprises 200 queries and 40,537 unique documents from the Web Track 2009-2012 dataset. Notably, queries #95 and #100 are excluded from our experiments due to the absence of diversity judgments. The remaining 198 queries consist of 3 to 8 manually annotated user intents with binary relevance ratings assigned at the intent level. To align with the TREC Web Track and prior approaches [32, 39], we adopt the top 50 results from Lemur as the prior relevance ranking.[2]

For the evaluation metrics, we adopt the official diversity evaluation metrics of Web Track, including ERR-IA [5], α-nDCG [7], and NRBP [8] to align with existing methods [14, 22]. These metrics assess the diversity of document rankings by explicitly rewarding novelty while penalizing redundancy. These evaluation metrics are all computed based on the top 20 ranking results. For significance testing, consistent with previous studies [22, 32, 39, 54], we conduct a two-tailed paired t-test with p-value < 0.05.

[1] ClueWeb09 Dataset: https://lemurproject.org/clueweb09.php/
[2] Lemur Service: http://boston.lti.cs.cmu.edu/Services/cluweb09_batch/

4.2 Baselines

We compare the proposed search agent with two types of baselines, including the ad-hoc search and the diversified search. The diversified search baselines can be roughly categorized into unsupervised methods and supervised methods.

Ad-hoc Search. `Lemur` and `ListMLE` [46] are two representative ad-hoc search methods without considering diversity. `Lemur` is the search model based on the Indri engine.

Diversified Search. (i) `xQuAD` [35] and `PM2` [9] are typical unsupervised diversification methods that both use a parameter λ to combine the relevance score and the diversity score of a document. Following `xQuAD` and `PM2`, `TxQuAD/TPM2` [10] further uses terms to model intents while `HxQuAD/HPM2` [17] introduces hierarchical subtopics with an additional parameter α.

(ii) `R-LTR` [54], `PAMM` [47], and `NTN` [48] are representative supervised methods using neural networks. We implement `NTN` based on `R-LTR` and `PAMM`. `DSSA` [20], `DVGAN` [22], `DESA` [32], `GDESA` [33], and `HAD` are explicit supervised models that use query suggestions as a proxy for actual user intents. We adopt the list-pairwise loss proposed by `DSSA` for training these models.

(iii) `DALETOR` [50], `Graph4DIV` [39], and `PAD` are implicit supervised search result diversification models. We implement the diversification-aware loss in `DALETOR` based on the evaluation metric α-nDCG [49]. `KEDIV` [42] introduces entities and their relationships from an external knowledge base to model the diversity of documents, while `CL4DIV` [14] devises contrastive learning tasks for initial document representation optimization.

(iv) We also include two latest closed-source LLMs `Claude3.5` (Claude 3.5 Sonnet) [1] and `GPT-4o` [29] for comparison and instruct them with specific prompts tailored for search result diversification. Due to the outstanding capabilities of recent reasoning models, we also introduce one of the most advanced reasoning LLMs, `Claude3.7` (Claude 3.7 Sonnet), for evaluation.

4.3 Implementation Details

For all supervised baseline models, we train the model with the full dataset and use five-fold cross-validation based on α-nDCG to select the best model. In contrast, our search agent operates without training data. The LLM is accessed through Anthropic and OpenAI API, including the Claude 3.5 Sonnet, Claude 3.7 Sonnet, and GPT-4o variants. We set the temperature of generation to 0.3 to balance uncertainty and variability in responses. A one-shot in-context example is employed in all instruction prompts. We set K in the greedy-based ranker to 20. For the LLMs without explicit reasoning ability (Claude 3.5 Sonnet and GPT-4o), we prompt them with three different prompts. Conversely, for the reasoning LLMs, such as Claude 3.7 Sonnet, we concatenate all prompts and ask the LLM to reason the whole process and generate the diversified document ranking with one prompt. The average cost of each query is around 0.05 dollars. For the specialization distillation, we use LLaMA 3.1-8B with a zero-shot generation instruction and evaluate the distilled model with five-fold cross-validation. Our code and more prompts and implementation details are released at Github.[3]

[3]Open-source code of our search agent `DIVAgent`: https://github.com/DengZhirui/DIVAgent/tree/main

Table 1: Overall performance of all methods. Zero-shot indicates model performance without task-specific training. The best zero-shot results are in bold. † indicates the model significantly outperforms zero-shot baselines with paired t-tests at p-value < 0.05 level.

Task	Method	Zero-Shot	ERR-IA	α-nDCG	NRBP
Ad-hoc search	`Lemur`	✓	.271	.369	.232
	`ListMLE`	✓	.287	.387	.249
	`Claude3.7`	✓	.339	.437	.308
Div. search	`R-LTR`	-	.303	.403	.267
	`PAMM`	-	.309	.411	.271
	`R-LTR-NTN`	-	.312	.415	.275
	`PAMM-NTN`	-	.311	.417	.272
	`DSSA`	-	.356	.456	.326
	`DALETOR`	-	.364	.461	.333
	`DESA`	-	.363	.464	.332
	`DVGAN`	-	.367	.465	.334
	`GDESA`	-	.369	.469	.337
	`Graph4DIV`	-	.370	.468	.338
	`HAD`	-	.387	.480	.361
	`PAD`	-	.386	.482	.357
	`KEDIV`	-	.390	.485	.362
	`CL4DIV`	-	.393	.486	.364
	`xQuAD`	✓	.317	.413	.284
	`TxQuAD`	✓	.308	.410	.272
	`HxQuAD`	✓	.326	.421	.294
	`PM2`	✓	.306	.411	.267
	`TPM2`	✓	.291	.399	.250
	`HPM2`	✓	.317	.420	.279
	`GPT-4o`	✓	.313	.410	.279
	`Claude3.5`	✓	.337	.435	.305
	`Claude3.7`	✓	.350	.447	.318
	`DIVAgent`	✓	**.386**†	**.478**†	**.358**†

4.4 Overall Results

The overall results of our proposed method `DIVAgent` and all baselines are shown in Table 1. We can find that:

(1) `DIVAgent` significantly outperforms existing ad-hoc and unsupervised diversified search methods on all evaluation metrics. Compared with the best LLM-based unsupervised baseline `Claude3.7`, `DIVAgent` achieves a significant edge with the absolute value of α-nDCG improved by 3.1%. This indicates that the superiority of our method stems from our search agent workflow rather than merely using LLMs. Besides, `DIVAgent` also achieves a remarkable performance improvement in comparison with `HxQuAD/HPM2`, which is the best unsupervised baseline without LLMs. `HxQuAD/HPM2` adopts a hierarchical structure to represent user intents and scores documents with manually designed functions, necessitating meticulous hyperparameter tuning. In contrast, our proposed diverse search agent `DIVAgent` mimics human cognitive processes during the search result diversification task. This approach requires only the provision of queries and candidate documents to produce diversified ranking results, significantly reducing the reliance on manually designed ranking functions and hyperparameter adjustments.

Table 2: The case study of DIVAgent on the query #50 "Dog Heat". Contents with the same color indicate the same topic.

Query #50	Dog Heat
User intent identification	[A] Understanding dog heat cycle, [B] Knowing the symptoms of dog heat, [C] Determining the ideal temperature for dogs, . . ., [E] Dog heatstroke prevention
Direct content extraction/ Core content compression	[1]: Female dogs have a 6-7 month heat cycle, with 6-8 days of receptivity to males. . . . [2]: Pet Street Mall offers a wide variety of dog beds in different styles, sizes, and materials, ... [49]: A female dog's heat cycle includes proestrus, estrus, and diestrus. Signs of heat include vaginal discharge, restlessness, and increased urination. . . . [50]: Dogs can suffer from heat stroke, especially in hot, humid weather. The symptoms and treatment of dog heat stroke . . .
Document intent alignment	I will identify the main topics covered in the documents based on the user intents: [1]. Understanding dog heat cycle. . . . [49]: Knowing the symptoms of dog heat. [50]: Dog heatstroke prevention.
Intent-guided ranker	I'll rerank based on relevance and topic diversity: [1]:[A] > [49]:[B] > [50]:[C] > . . .
Final ranking	[1] [49] [50] . . .

(2) Intriguingly, DIVAgent can achieve competitive results, and even surpass some supervised methods, particularly in terms of ERR-IA and NRBP. This narrowing of the disparity between unsupervised and supervised models is attributable to the intent-aware reasoning process and the intent-guided output format of DIVAgent, as well as the textual information modeling capabilities of LLMs. By employing prompts designed for diversification, DIVAgent eliminates the reliance on large-scale annotated data while maintaining comparable effectiveness. Moreover, through prompt-based interactions with LLMs, DIVAgent naturally generates a transparent decision-making process and explainable diversified document ranking. This interpretability distinguishes DIVAgent from traditional black-box models [14, 32], facilitating more accountable and user-trustworthy search result diversification systems.

4.5 Case Study

To provide an intuitive demonstration of our DIVAgent, we first conduct a case study to observe the role of each component in our workflow. We randomly select query #50 *Dog Heat* and illustrate its related intermediate results in Table 2.

Initially, when DIVAgent receives the query *Dog Heat*, the intent-aware module first mirrors human search behaviors to analyze the input and predict several potential user intents. Concurrently, our search agent will go through each candidate document and extract its core content with direct content extraction or a core content compression strategy. By mimicking the human process of selecting diverse information, our search agent performs document-intent alignment, reasoning about the intent coverage of each document. To facilitate the generation of a diverse document ranking, the intent-guided ranker integrates a greedy-based selection strategy with an intent-guided output format. As illustrated in Table 2, DIVAgent can successfully derive a document ranking encompassing distinct topics at the top of the ranking list, indicating the benefit of mimicking human search processes for developing DIVAgent.

Table 3: Results of ablation studies with different components belonging to three modules, respectively.

Method	ERR-IA	α-nDCG	NRBP
DIVAgent	**.386**	**.478**	**.358**
w/o User intent identification	.370	.464	.339
w/o Document-intent alignment	.376	.470	.347
w/o Intent-guided ranking	.366	.460	.336
w/ Direct content extraction	.386	.478	.358
w/ Core content compression	.384	.479	.357

4.6 Ablation Studies

Next, we conduct ablation studies to explore the influence of different components in DIVAgent. We also depict the performance of different document content modeling mechanisms.

The results are presented in Table 3, and we can observe that removing any individual component results in a noticeable decline in performance. This finding demonstrates the importance of each component in contributing to an effective diversified search agent. Meanwhile, the exclusion of the intent-guided ranking strategy leads to the most severe decrease in performance. This indicates that explicitly guiding the agent to output the document identifier with its covered intent identifier during the document ranking process can facilitate the overall diversity. Besides, both user intent identification and document-intent alignment contribute a lot to overall performance. Eliminating either component causes a considerable drop in all metrics (*e.g.*, ERR_IA: 0.386 $\rightarrow$ 0.370/0.376 and α-nDCG: 0.478 $\rightarrow$ 0.464/0.470). This observation is consistent with our assumption, as more precise and comprehensive modeling of user intent and document intent coverage judgment encourages the measurement of the documents' novelty. Moreover, we observe that directly extracting document content yields comparable performance to compressing it into core summaries. A possible reason is that important information on web pages is typically front-loaded, allowing the initial segment of each document to effectively capture its core content without additional summarization. Given that content compression introduces additional computational cost, our search agent adopts direct content extraction for document content modeling, offering a more efficient yet equally effective solution.

4.7 Piecewise Evaluation

In this section, we conduct a piecewise evaluation to explore the performance of each component in our search agent. To evaluate the quality of user intents identified by our search agent, we directly compare the generated intent sets against those annotated in the TREC Web Track datasets. Given that exact matching may overlook semantic similarities and lead to overly rigid evaluation results, we adopt GPT-4 to assess the relative completeness, and distinctiveness of the two intent sets. For document-intent alignment, direct comparison is challenging due to the different intent taxonomies between our method and the annotations. To address

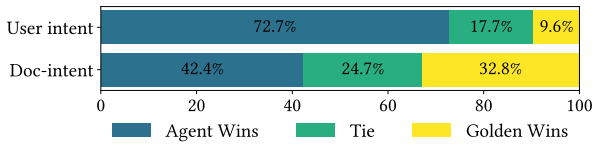

Figure 3: Piecewise evaluation results for user intent identification and document-intent alignment.

this, we propose to separately use both the user intents and corresponding document-intent relevance annotations from the TREC dataset and our search agent, and prompt an LLM for diversified document ranking. The α-nDCG is compared for evaluation. Results are presented in Figure 3.

Compared to the manually labeled user intents in the TREC Web Track 2009–2012 dataset, our search agent achieves superior performance, outperforming the TREC annotations in 72.5% of the evaluated cases. The TREC user intents are constructed from commercial query logs, which, while effective, are inherently constrained by historical user behavior and high-frequency, mainstream intents. In contrast, our search agent uses LLMs to directly generate user intents in a few-shot setting. This enables the discovery of long-tail user intents and eliminates the need for human intent construction.

As for document-intent alignment, the rankings generated based on our search agent's judgments slightly outperform those derived from the human-labeled annotations. We attribute this improvement to the agent's ability to capture subtle intent-document relationships that may be overlooked by log-derived taxonomies. This indicates that even without manual supervision, our search agent can support high-quality, diversified document ranking.

4.8 Influence of the Backbone Model

To investigate the influence of the backbone model on the performance of `DIVAgent`, we conduct experiments with three types of models: (i) open-sourced models (*i.e.*, Llama 3.1-8B and Llama 3.1-70B); (ii) closed-sourced models without deep reasoning (*i.e.*, GPT-4o and Claude 3.5 sonnet); (iii) closed-sourced models with deep reasoning (*i.e.*, Claude 3.7 sonnet and Gemini-2.5 pro). The results are shown in Figure 4.

First, we can observe that `DIVAgent` achieves consistent and robust efficacy across models with various parameter scales. Notably, even with a smaller backbone model, such as Llama 3.1-8B, our search agent can still outperform existing baselines in few-shot scenarios. This performance stability demonstrates the adaptability of our search agent to different configurations, guaranteeing its reliability in achieving effective results even in resource-constrained environments. Second, in contrast to both open-source and closed-source LLMs that lack explicit reasoning mechanisms, our search agent, when powered by a reasoning-capable model (*e.g.*, Claude 3.7 Sonnet), demonstrates superior performance. This advancement indicates that the reasoning capability of the backbone model enables a comprehensive and step-by-step analysis of user intent and document intent coverage, resulting in better and more explainable diversified rankings.

4.9 Parameter Sensitivity

In `DIVAgent`, the number of user intents identified in PROMPT 1 and the length of the extracted document content in Section 3.2 are

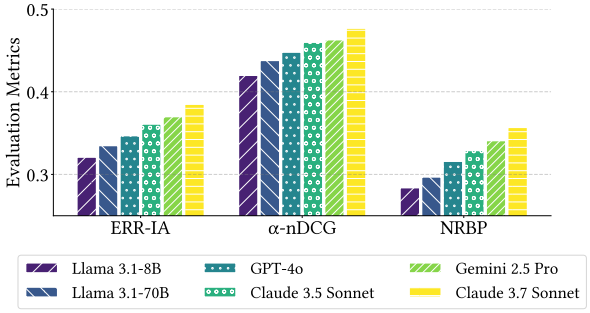

Figure 4: Performance with different backbone models on all evaluation metrics.

two important hyperparameters affecting the performance of our search agent. To investigate their impact, we conduct experiments with different hyperparameter settings and record the performance. The results are shown in Figure 5.

Number of Intents. To identify the optimal number of user intents for the prompt, we increase it from zero to 20 with equal spacing in steps of 5, while closely monitoring the performance changes in terms of all evaluation metrics. Zero means no explicit mention of the limit on the number of generated user intents. As depicted in Figure 5 (a), the performance improves progressively as the window size increases from zero to 10. The peak of the performance is reached at the window size set to 10. When the window size exceeds 10, the performance starts to degrade. This phenomenon can be attributed to the fact that as the number becomes larger, the model tends to generate overly fragmented or redundant intents, which not only introduces semantic overlaps among intents but also disrupts the latter ranking. Therefore, carefully selecting an appropriate user intent number is crucial for balancing context comprehension and diversification quality.

Document Length. Secondly, we investigate the impact of document input length on overall ranking performance. We test the number of input tokens per document from 50 to 250 and report the results in Figure 5 (b). The findings reveal that the performance can be gradually improved as the length increases from 50 to 100. This implies that providing additional context enables the model to better understand document relevance and intent coverage. However, when the document length exceeds 100, further increases do not yield additional performance gains. This phenomenon may be attributed to two factors. First, webpages exhibit a front-loaded structure, where the most relevant or informative content appears at the beginning. As such, extending the input length beyond a certain threshold adds less informative or redundant content. Second, LLMs are known to suffer from the "lost-in-the-middle" problem [23], where important information located in the middle of a long sequence receives insufficient attention. These issues ultimately constrain the benefit of long document inputs.

4.10 Analysis of Efficiency

In addition to effectiveness, efficiency is also a critical metric for determining user satisfaction. In this section, we break down efficiency into two primary dimensions: training efficiency and inference efficiency. Training efficiency indicates the duration required to train

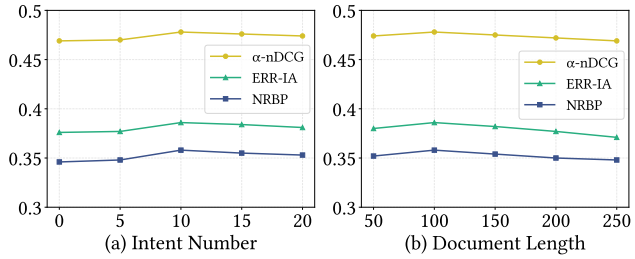

Figure 5: Performance with different hyperparameters.

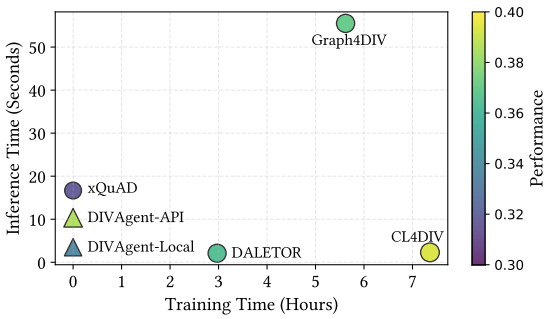

Figure 6: Analysis of training and inference efficiency on different search result diversification models. The triangles represent DIVAgent while circles represent baselines.

the model on the labeled dataset, whereas inference efficiency refers to the query latency during practical application, typically quantified by seconds. To provide a thorough evaluation of the inference efficiency of our diversified search agent DIVAgent, we conduct experiments with both API invocation (Claude 3.7 sonnet) and local deployment (Llama 3.1-70B).

As depicted in Figure 6, our proposed search agent DIVAgent demonstrates a balance between efficiency and effectiveness. In contrast to supervised methods (*e.g.*, DALETOR and CL4DIV), which require several hours of training on labeled datasets to achieve optimal performance, DIVAgent performs effectively without task-specific training. By using the inherent capabilities of LLMs, DIVAgent eliminates the need for time-consuming training processes. Furthermore, compared to unsupervised models (*e.g.*, xQuAD), DIVAgent demonstrates superior inference efficiency, particularly in local deployment scenarios. Our experiments indicate that the local deployment of DIVAgent substantially alleviates network latency issues, resulting in faster query processing times.

These efficiency advantages position DIVAgent as a practical and scalable solution for real-world applications, particularly in scenarios characterized by a scarcity of human labeled data. This approach delivers competitive performance without the need for additional training.

4.11 Performance of Specialization Distillation

As illustrated in Section 3.5, we propose a specialization distillation strategy to transfer the diversified ranking ability of powerful expert LLMs into a specialized smaller model. In this section, we conduct experiments with Claude 3.7 Sonnet and LLaMA 3.1-8B, and the results are summarized in Table 4.

Table 4: Results of specialization distillation.

Model	Label	ERR-IA	α-nDCG	NRBP
Claude 3.7 Sonnet	-	**.386**	**.478**	**.358**
LLaMA 3.1-8B	-	.322	.421	.285
LLaMA 3.1-8B	Claude 3.7 Sonnet	.368	.468	.336

From the results, we can find that the student model distilled from Claude 3.7 Sonnet achieves performance close to the expert across all evaluation metrics. Notably, compared to the original version, the distilled model exhibits significant performance improvements. These findings demonstrate the effectiveness of using LLM-generated signals for ranking specialization.

5 Conclusion and Future Work

Search result diversification plays an important role in improving user satisfaction. Previous supervised methods, while effective but typically rely on massive training data. Conversely, unsupervised approaches eliminate the need for training but require heuristically constructed ranking functions. In this paper, we combine the advantages of supervised methods and unsupervised methods to achieve effective search result diversification without heavy human involvement. Drawing inspiration from human search processes, we introduce a diversified search agent DIVAgent that imitates the process of human search and finds diverse information. The proposed search agent incorporates three essential modules tailored for search result diversification: an intent-aware module, a memory module, and an intent-guided ranker module. For each query with its corresponding document list, the intent-aware module initially predicts potential user intents and analyzes each document's content for document-intent alignment. Information beneficial to the ongoing task is stored in the memory module for further processing. Finally, the intent-guided ranker generates the intent-guided output with a greedy-based selection strategy, which explicitly indicates the intent coverage of each document. Experimental results demonstrate the effectiveness of our diversified search agent DIVAgent even without task-specific fine-tuning.

In future work, we plan to integrate DIVAgent with a generator to facilitate direct responses to user queries.

Acknowledgments

Zhicheng Dou is the corresponding author. This work was supported by National Science and Technology Major Project No. 2022ZD0120103, National Natural Science Foundation of China No. 62272467, by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No 101070212. The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

GenAI Usage Disclosure

We have used generative AI software tools to polish and improve the clarity of our writing.

References

[1] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://api.semanticscholar.org/CorpusID:268232499

[2] Lawrence W Barsalou. 2014. *Cognitive Psychology: An Overview for Cognitive Scientists*. Psychology Press.

[3] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. The ClueWeb09 dataset, 2009. http://boston.lti.cs.cmu.edu/Data/clueweb09.

[4] Jaime G. Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM, 335–336. https://doi.org/10.1145/290941.291025

[5] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin (Eds.). ACM, 621–630. https://doi.org/10.1145/1645953.1646033

[6] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. 2024. MindSearch: Mimicking Human Minds Elicits Deep AI Searcher. *arXiv preprint arXiv:2407.20183* (2024).

[7] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 659–666. https://doi.org/10.1145/1390334.1390446

[8] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Advances in Information Retrieval Theory, Second International Conference on the Theory of Information Retrieval, ICTIR 2009, Cambridge, UK, September 10-12, 2009, Proceedings (Lecture Notes in Computer Science, Vol. 5766)*, Leif Azzopardi, Gabriella Kazai, Stephen E. Robertson, Stefan M. Rüger, Milad Shokouhi, Dawei Song, and Emine Yilmaz (Eds.). Springer, 188–199. https://doi.org/10.1007/978-3-642-04417-5_17

[9] Van Dang and W. Bruce Croft. 2012. Diversity by Proportionality: An Election-based Approach to Search Result Diversification. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson (Eds.). ACM, 65–74. https://doi.org/10.1145/2348283.2348296

[10] Van Dang and W. Bruce Croft. 2013. Term Level Search Result Diversification. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai (Eds.). ACM, 603–612. https://doi.org/10.1145/2484028.2484095

[11] Zhirui Deng, Zhicheng Dou, Zhan Su, and Ji-Rong Wen. 2024. Multi-grained Document Modeling for Search Result Diversification. *ACM Trans. Inf. Syst.* 42, 5 (2024), 126:1–126:22. https://doi.org/10.1145/3652852

[12] Zhirui Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. DeepQFM: a deep learning based query facets mining method. *Information Retrieval Journal* 26, 1 (2023), 9.

[13] Zhirui Deng, Zhicheng Dou, Yutao Zhu, Xubo Qin, Pengchao Cheng, Jiangxu Wu, and Hao Wang. 2024. JDivPS: A Diversified Product Search Dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1152–1161.

[14] Zhirui Deng, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2024. CL4DIV: A Contrastive Learning Framework for Search Result Diversification. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 171–180.

[15] Zhirui Deng, Zhicheng Dou, Yutao Zhu, Ji-Rong Wen, Ruibin Xiong, Mang Wang, and Weipeng Chen. 2024. From Novice to Expert: LLM Agent Policy Optimization via Step-wise Reinforcement Learning. arXiv:2411.03817 [cs.AI] https://arxiv.org/abs/2411.03817

[16] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A Large-scale Evaluation and Analysis of Personalized Search strategies. In *Proceedings of the 16th international conference on World Wide Web*. 581–590.

[17] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search Result Diversification Based on Hierarchical Intents. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 63–72. https://doi.org/10.1145/2806416.2806455

[18] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Inf. Process. Manag.* 36, 2 (2000), 207–227. https://doi.org/10.1016/S0306-4573(99)00056-4

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[20] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. 2017. Learning to Diversify Search Results via Subtopic Attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 545–554. https://doi.org/10.1145/3077136.3080805

[21] Nayoung Kim and Chang S Nam. 2020. Adaptive Control of Thought-rational (ACT-R): Applying a Cognitive Architecture to Neuroergonomics. *Neuroergonomics: Principles and Practice* (2020), 105–114.

[22] Jiongnan Liu, Zhicheng Dou, Xiaojie Wang, Shuqi Lu, and Ji-Rong Wen. 2020. DVGAN: A Minimax Game for Search Result Diversification Combining Explicit and Implicit Features. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 479–488. https://doi.org/10.1145/3397271.3401084

[23] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[24] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-Shot Listwise Document Reranking with a Large Language Model. *CoRR* abs/2305.02156 (2023). https://doi.org/10.48550/ARXIV.2305.02156 arXiv:2305.02156

[25] David Meunier, Renaud Lambiotte, Alex Fornito, Karen Ersche, and Edward T Bullmore. 2009. Hierarchical Modularity in Human Brain Functional Networks. *Frontiers in neuroinformatics* 3 (2009), 571.

[26] Microsoft. 2023. Confirmed: The New Bing Runs on OpenAI's GPT-4. https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAIâĂŹs-GPT-4

[27] John Morkes and Jakob Nielsen. 1997. Concise, scannable, and objective: How to write for the Web. *Useit. com* 51, 1 (1997), 1–17.

[28] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted Question-answering with Human Feedback. *arXiv preprint arXiv:2112.09332* (2021).

[29] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[30] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *CoRR* abs/2309.15088 (2023). https://doi.org/10.48550/ARXIV.2309.15088 arXiv:2309.15088

[31] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *CoRR* abs/2312.02724 (2023). https://doi.org/10.48550/ARXIV.2312.02724 arXiv:2312.02724

[32] Xubo Qin, Zhicheng Dou, and Ji-Rong Wen. 2020. Diversifying Search Results using Self-Attention Network. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1265–1274. https://doi.org/10.1145/3340531.3411914

[33] Xubo Qin, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2022. GDESA: Greedy Diversity Encoder with Self-Attention for Search Results Diversification. *ACM Transactions on Information Systems (TOIS)* (2022).

[34] Clara Rus, Jasmin Kareem, Chen Xu, Yuanna Liu, Zhirui Deng, and Maria Heuss. 2025. AMS42 at the NTCIR-18 FairWeb-2 Task. *Proceedings of NTCIR-18* (2025).

[35] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting Query Reformulations for Web Search Result Diversification. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti (Eds.). ACM, 881–890. https://doi.org/10.1145/1772690.1772780

[36] Craig Silverstein, Monika Rauch Henzinger, Hannes Marais, and Michael Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (1999), 6–12. https://doi.org/10.1145/331403.331405

[37] Robert L Solso, M Kimberly MacLin, and Otto H MacLin. 2005. *Cognitive psychology*. Pearson Education New Zealand.

[38] Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying Ambiguous Queries in Web Search. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy (Eds.). ACM, 1169–1170. https://doi.org/10.1145/1242572.1242749

[39] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen. 2021. Modeling Intent Graph for Search Result Diversification. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 736–746. https:

//doi.org/10.1145/3404835.3462872

[40] Zhan Su, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2022. Knowledge Enhanced Search Result Diversification. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 1687–1695. https://doi.org/10.1145/3534678.3539459

[41] Zhan Su, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2024. Passage-aware Search Result Diversification. *ACM Trans. Inf. Syst.* 42, 5 (2024), 136:1–136:29. https://doi.org/10.1145/3653672

[42] Zhan Su, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2022. Knowledge Enhanced Search Result Diversification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1687–1695.

[43] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 14918–14937. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.923

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[45] TREC. 2009. TREC 2009 Web Track. https://trec.nist.gov/data/web09.html

[46] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise Approach to Learning to Rank: Theory and Algorithm. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008 (ACM International Conference Proceeding Series, Vol. 307)*, William W. Cohen, Andrew McCallum, and Sam T. Roweis (Eds.). ACM, 1192–1199. https://doi.org/10.1145/1390156.1390306

[47] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning Maximal Marginal Relevance Model via Directly Optimizing Diversity Evaluation Measures. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 113–122. https://doi.org/10.1145/2766462.2767710

[48] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling Document Novelty with Neural Tensor Network for Search Result Diversification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 395–404. https://doi.org/10.1145/2911451.2911498

[49] Chen Xu, Zhirui Deng, Clara Rus, Xiaopeng Ye, Yuanna Liu, Jun Xu, Zhicheng Dou, Ji-Rong Wen, and Maarten de Rijke. 2025. FairDiverse: A Comprehensive Toolkit for Fairness-and Diversity-aware Information Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3540–3550.

[50] Le Yan, Zhen Qin, Rama Kumar Pasumarthi, Xuanhui Wang, and Michael Bendersky. 2021. Diversification-Aware Learning to Rank using Distributed Representation. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 127–136. https://doi.org/10.1145/3442381.3449831

[51] Puxuan Yu, Razieh Rahimi, Zhiqi Huang, and James Allan. 2023. Search Result Diversification Using Query Aspects as Bottlenecks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 3040–3051. https://doi.org/10.1145/3583780.3615050

[52] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than Retrieve: Large Language Models are Strong Context Generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=fB0hRu9GZUS

[53] Yisong Yue and Thorsten Joachims. 2008. Predicting Diverse Subsets using Structural SVMs. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008 (ACM International Conference Proceeding Series, Vol. 307)*, William W. Cohen, Andrew McCallum, and Sam T. Roweis (Eds.). ACM, 1224–1231. https://doi.org/10.1145/1390156.1390310

[54] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for Search Result Diversification. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin (Eds.). ACM, 293–302. https://doi.org/10.1145/2600428.2609634