A Pointwise Method for Search Result Diversification via Intent Matching

Chunyang Li
Xidian University
Xi'an, China
0009-0000-3684-2849

Zhichao Huang *
JD Intelligent Cities Research
Beijing, China
0000-0002-7662-5184

Xubo Qin *
Renmin University of China
Beijing, China
0000-0002-3717-9492

***Abstract*—Search result diversification aims to mitigate query ambiguity by presenting documents that collectively satisfy multiple user intents. However, many supervised diversification models focus on listwise interaction modeling while implicitly assuming that the initial candidates already provide high-quality intent signals; in practice, weak intent matching introduces noise that limits the effectiveness of even sophisticated diverse rankers. We propose Intent Matching Enhanced Diversification (IMEDiv), a decoupled two-stage framework that explicitly separates *intent matching* from *diverse re-ranking*. In the first stage, we fine-tune an instruction-following, decoder-only LLM as a pointwise intent matcher to score each document by its intent coverage given a query and its subtopics, producing an intent-enhanced ranking list. In the second stage, a lightweight subtopic-aware ranker reduces redundancy on top of these intent-enhanced candidates. Importantly, this stage can be replaced by existing diverse ranking models, making IMEDiv a plug-and-play relevance module. Experiments on two public datasets demonstrate that strengthening intent matching alone is highly effective (IMEDiv (Intent) improves α-nDCG@20 from 0.452 to 0.467 over DSSA), and combining it with diverse re-ranking further improves diversification (IMEDiv (Intent+Div) reaches α-nDCG@20 0.476). Detailed analyses show that the gains primarily come from improving intent coverage at top ranks, highlighting the overlooked but foundational role of intent matching in diversification.**

***Index Terms*—Search Result Diversification, Subtopic, Intent Matching, Large Language Models**

I. Introduction

In search engine research, studies have consistently shown users' preference for short queries. However, such brevity often leads to vague or ambiguous search intentions [1], [2]. For instance, the query "Java" could refer to either the "Java programming language" or the "JAVA island". Even for the same user, search needs may cover multiple aspects, from downloading software to finding tutorials and IDE recommendations. Search result diversification addresses this challenge by optimizing document rankings. The process involves two key steps: Firstly, initial document ranking based on relevance scores. Then, re-ranking through diversification models to enhance the variety. This approach enables users with different information needs to efficiently locate relevant content within a single search session. By presenting diverse perspectives in the results list, the system better accommodates various interpretations of ambiguous queries.

*: Joint corresponding authors (iceshzc@gmail.com, qratosone@live.com)

Diversification methodologies are generally categorized as implicit or explicit methods. Implicit methods analyze document interaction patterns to quantify novelty via pairwise dissimilarity. Explicit methods, in contrast, use a **subtopic-centric** framework to directly optimize for comprehensive subtopic coverage. Recent ensemble approaches combine both, integrating document interaction modeling and subtopic coverage. Subtopic-oriented methodologies typically use a two-stage architecture, as depicted in Figure 1: an intent matching component to evaluate document-subtopic relevance during initial ranking, and a diverse ranking component to dynamically adjust document priorities based on coverage.

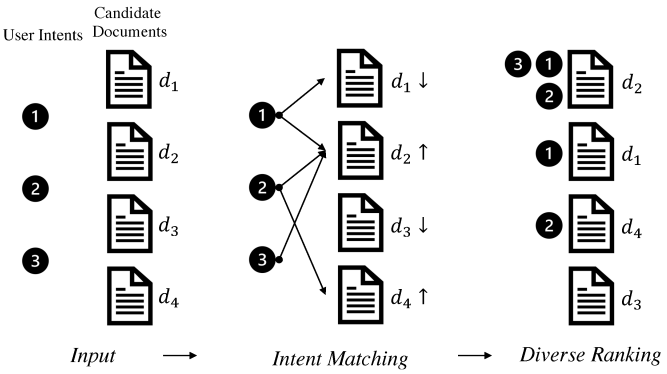

Fig. 1: Two-stage framework: intent matching and diverse ranking.

However, current research predominantly **emphasizes second-stage document interaction modeling through listwise loss functions, with insufficient attention to first-stage intent matching**. Prevailing methodologies operate under the assumption that initial document inputs sufficiently encapsulate user intents, thereby reducing the task to re-ranking optimization. To enhance diverse ranking, researchers have used complex neural architectures, including attention-based RNNs [3], GANs [4], and self-attention encoders [5] to compute relevance scores with traditional features. Taking DSSA [3] as a representative explicit method, it computes relevance scores through conventional features combined with distributed representations of queries/subtopics and documents. Nevertheless, data scarcity necessitates reliance on unsupervised representation learning techniques like doc2vec [6],

which yield suboptimal semantic embeddings as evidenced by [7]. This dual challenge of inadequate training data and imperfect representations consequently complicates both training data sampling and model training.

To address these, we present two key arguments challenging in search result diversification. First, we **emphasize the critical and foundational role of intent matching in diversification tasks**, while existing approaches frequently overlook its importance. Although the goal of search result diversification is to satisfy diverse user intents, presenting "diversified" yet irrelevant documents fails to fulfill any actual information needs. In contrast, even redundant documents that effectively match user intents inherently offer greater utility than irrelevant results. Second, prior work typically combines intent matching and diverse ranking into a single-stage process through end-to-end optimization. We propose that **the intent matching component can be independently optimized** as a pointwise ranking task under the Probability Ranking Principle (PRP) [8]. In our context, PRP motivates treating a document's *intent-satisfaction propensity* as primarily determined by its own matching signal with the query (and subtopics), rather than by explicit dependencies on other candidates.

In this paper, we propose a novel two-stage diversification framework that explicitly separates intent matching from diverse ranking. While previous research has extensively studied diverse ranking components, we primarily focus on advancing intent matching through modern large language models (LLMs). Our framework's first stage employs an LLM-powered intent-matching component that operates as an independent pointwise reranker. This "intent-enhanced re-ranking" mechanism prioritizes documents addressing multiple user intents, ensuring broader intent coverage in higher ranking positions. The second stage incorporates a streamlined, diverse ranking component utilizing subtopic-aware features and a max-pooling selection strategy. This lightweight module further refines the intent-enhanced rankings by reducing redundancy while maintaining intent coverage. Crucially, our architecture permits direct substitution of this component with existing SOTA diverse ranking models, enabling enhanced result diversity and improved intent satisfaction. Experimental results demonstrate that our intent-matching component achieves competitive performance comparable to sophisticated listwise diversification methods. Moreover, applying diverse ranking components to our intent-enhanced outputs yields significantly improved diversification while preserving intent coverage. The main contributions are listed as follows:

- We propose a decoupled two-stage diversification framework that explicitly separates intent matching from diverse re-ranking, enabling independent optimization of intent matching as a PRP-style pointwise ranking task.
- We introduce an instruction-tuned LLM intent matcher that scores each document by its intent coverage given a query and its subtopics, providing higher-quality intent-matching signals without handcrafted features.
- Extensive experiments on public datasets demonstrate that strengthening intent matching can significantly improve diversification effectiveness, and the proposed intent matcher can serve as a plug-and-play component to enhance existing diversification frameworks.

II. Related Work

In this section, we briefly review the previous works on search result diversification and the additivity of the diversification approaches over a relevance baseline.

A. Search Result Diversification Methods

Search result diversification methods can be categorized into implicit, explicit, and ensemble approaches. Early implicit methods, like MMR [9], greedily select documents based on query relevance and document novelty (measured via similarity), using handcrafted features. Extensions like SVM-DIV and DALETOR automate feature learning. Recent implicit methods utilize advanced neural architectures, such as graph convolutional networks [10] and multi-view fusion techniques [11], to judge diversity via user intent coverage rather than document similarity (e.g., Graph4Div [12], CL4DIV [13]). Explicit approaches (e.g., xQuAD [14], PM2 [15], DSSA [3]) model novelty as coverage of new user intents (subtopics). DSSA is supervised, employing RNNs and attention mechanisms to model subtopic distributions. However, it relies on weak relevance features and unsupervised doc2vec [6] embeddings, neglecting semantic subtopic signals. Furthermore, its ranking function is a simplistic linear layer rather than a well-designed learning-to-rank component [7], leading to suboptimal intent matching. DVGAN [4] and DESA [5] inherit these limitations as they also depend on these weak relevance features for subtopic modeling. Ensemble methods (e.g., DVGAN, DESA [5]) combine implicit and explicit features, but also suffer from suboptimal relevance matching due to reliance on weak features. Yigit-Sert et al. [16] proposed learning-to-rank methods for subtopic matching, but their approach depends on unavailable ideal user intents, limiting practical applicability.

B. Neural and Large Language Model based Ranking

Recently, with the achievement of neural-based NLP methods, researchers have proposed a group of language-based relevance matching models, e.g., BERT [17], [18], to model the semantical intent. The vanilla BERT-based relevance matching model takes the query and document tokens concatenated with a special token `[SEP]` as input, and the output vector corresponding to the `[CLS]` token is used as the representation of the query-document pair. Furthermore, large language models (LLMs) [19], [20] with billions of parameters have also been adapted to the task of search and ranking. Some methods, e.g., LRL [21], RankGPT [22], and PRP [23] have explored pairwise or listwise prompting approaches for zero-shot reranking tasks. Other approaches, such as Reinforced Query Reasoners [24], focus on reasoning-intensive retrieval tasks to enhance query understanding. Additionally, retrieval-augmented generation frameworks [25], [26] have demonstrated effectiveness in summarizing data resources and processing complex information, while some works [27] have introduced graph

structure [28] to further increase the performance. In addition, some other approaches e.g. RankLLama [29] or bge-reranker-v2 [30] are based on fine-tuning a LLM decoder to function as a pointwise reranker specifically. However, to the best of our knowledge, none of these existing LLM-based approaches is used in search result diversification. In this work, we use an LLM-based component to capture intent-matching signals for each candidate document corresponding to the given query. The LLM is fine-tuned to follow the instructions for judging if a document has covered the intents behind the given query.

C. The Additivity Over Weak and Strong Baselines

In a real IR system, faced with a user-issued query, the system usually processes it with several stages, including retrieving, initial ranking, re-ranking, etc. The diversification works in the re-ranking stage. To investigate the additivity of diversification over different initial ranking baselines (e.g., the Indri search engine), Kharazmi et al. [31] presented a study for the effect of different baselines. Their study showed that for a diversification approach, the statistically significant improvement over non-diversified baselines may disappear when a weak baseline is replaced by a stronger one. In other words, with more sophisticated relevance scoring approaches being used, even the initial ranking could be diverse. Further studies [16] revealed that diversification methods should be carefully tuned to obtain further diversification effectiveness. Previous studies often pay attention to the relation of relevance and diversity in non-diversified initial ranking, but the effect of the intent matching component is rarely studied.

III. The Proposed Framework

A. Problem Formulation

In this paper, we split the search result diversification task into two stages: the intent matching stage and the diverse re-ranking stage. Since the diverse re-ranking stage has been widely explored by previous works, we only focus on the intent matching stage. The task of intent matching can be defined as a task of PRP-based ranking. Given a query q with a set of subtopics $\mathcal{I}$ and a list of candidate documents $\mathcal{D}$, the simplified diverse ranking task aims to return a new ranked document list $\mathcal{R}$. Here $\mathcal{D}$ is an initial relevance ranking list without diversification. For the reranked list $\mathcal{R}_{\text{intent}}$, we only consider the intent coverage of each candidate document. As an explicit approach, our framework models the intent coverage of each document with subtopics $\mathcal{I}$, and a diversified document should satisfy more intents at former ranking positions. The reranked list $\mathcal{R}_{\text{intent}}$ here will satisfy multiple user intents at former positions, while it may suffer from redundancy. With a diverse ranking approach, the former ranking positions in $\mathcal{R}_{\text{intent}}$ can be further reranked to improve diversity. A diverse ranker will take the top-N documents of $\mathcal{R}_{\text{intent}}$ as input, returning the diversified ranking list $\mathcal{R}_{\text{div}}$ to further reduce the redundancy.

B. LLM-Based Intent Matcher

Focusing on intent coverage of each document, we propose an intent matcher with large language models (LLMs), which is formulated as a pointwise "reranker" model. Our proposed approach involves the following inputs: a query and its corresponding subtopics, and a candidate document to judge. For each candidate document d, the ranking score s of d with the query q and the subtopics $\mathcal{I}$ are defined as follows:

$$s = \text{Intent}(q, \mathcal{I}, d), \tag{1}$$

where the ranking score s denotes an intent-coverage score for the document d with respect to q. Assuming that the document d_i covers more intents than another document d_j, it will lead to the score s_i being higher than s_j. Unlike previous works [3], [5], [13], our LLM-based intent-matcher only requires the raw tokens of q, $\mathcal{I}$, and d. The model is trained end-to-end with no extra features required.

This score is an *intent-coverage score* that indicates how strongly d is predicted to cover the intents behind q. We use a decoder-based LLM to judge the intent coverage. The input, $\langle q, \mathcal{I}, d \rangle$, is concatenated with a predefined instruction template. The intent coverage score is then calculated as:

$$\text{Input} = \text{Instruct}(\langle q, \mathcal{I}, d \rangle), \tag{2}$$

$$\text{Output} = \text{Decoder}(\text{Input})[-1], \tag{3}$$

$$\text{Logits} = \text{MLP}(\text{Output}), \tag{4}$$

$$s = \text{Logits}[\text{loc}_{\text{yes}}]. \tag{5}$$

Here, $\text{Instruct}(\cdot)$ is the prompt template, $\text{Decoder}(\cdot)$ is the LLM, and loc_{yes} selects the output logit for the "Yes" token, which serves as the score s. In practice, we use the logit as a ranking score (rather than a calibrated probability), which is sufficient for pointwise reranking. For implementation, loc_{yes} is the token id obtained from the model tokenizer for the string "Yes". The prompt details are in Fig. 2.

Instruction Templates for Intent-Matching

{query}\n
{passage}\n
Consider a query A and a passage B. The subtopics of this query A are as follows:
{subtopics (1), subtopics (2), ···}
These subtopics are specific and subdivided topics related to the query. Please determine whether passage B contains possible intents related to query A, and give the answer by predicting 'yes' or 'no'

Fig. 2: Prompt template for the LLM-based intent-matcher.

C. Max-Pooling Based Diverse Ranker

To implement the diverse ranker, we adopt a max-pooling based document selection strategy inspired by DSSA. Given $|\mathcal{I}|$ subtopics for query q, a subtopic $q_i \in \mathcal{I}$, and a document d, the feature-based exact matching score is:

$$s_{q,d} = \text{MLP}(x_{q,d}), \tag{6}$$

$$s_{q_i,d} = \text{MLP}(x_{q_i,d}), \quad i \in [1, |\mathcal{I}|], \tag{7}$$

where $x_{q,d}$ and $x_{q_i,d}$ are relevance feature vectors, and $\text{MLP}(x) = \text{ReLU}(x^\top w_1 + b_1)^\top w_2 + b_2$ is a two-layer linear perceptron. These scores are generated using traditional IR features like BM25, TF-IDF, and PageRank. For a selected

document sequence $\mathcal{C}$, the subtopic satisfaction state $S_{I,\mathcal{C}}$ is the concatenation of each subtopic score. A max-pooling score is then calculated to indicate the satisfaction of every subtopic for the whole sequence:

$$S(q_j, \mathcal{C}) = \max\big([s(q_j, d_1); \ldots; s(q_j, d_t)]\big), \tag{8}$$

where $j \in [1, m]$ and $t = |C|$. Then $S_{\mathcal{I},\mathcal{C}}$ produces a subtopic weight based on max-pooling with the softmax function, $M_{\mathcal{C}} = \mathrm{softmax}(-S_{\mathcal{I},\mathcal{C}})$. Here, the subtopic score of the selected sequence $S_{\mathcal{I},\mathcal{C}}$ measures the subtopic satisfaction of the sequence of previously selected documents, and a subtopic that has already been satisfied will be punished with a lower weight based on a negative softmax score. As a result, the framework will focus on selecting novel documents that cover subtopics, which has not been satisfied yet. Here, we use the max-pooling strategy to simplify the problem and address the effect of the intent-matching component in the diverse ranking task. It should be clarified that our framework is flexible and the diverse ranking component can be replaced with a more effective neural-based component, for example, the subtopic attention used in DSSA [3].

With the matching scores of the subtopics and the max-pooling based on subtopic weights, the overall diverse ranking scores can be defined as:

$$s_{\mathcal{I},\mathcal{C},d} = {S_{\mathcal{I},d}}^{\top} M_{\mathcal{C}}, \tag{9}$$

$$s_d = s_{q,d} w_r + s_{\mathcal{I},\mathcal{C},d} w_s. \tag{10}$$

For each ranking position, the model tries to select the best document covering the most unsatisfied subtopics. When a document is selected, it will be added to $\mathcal{C}$. As a result, the max-pooling weights $M_{\mathcal{C}}$ will change with the expansion of $\mathcal{C}$. In this work, we implement the diverse ranking system with the document selection strategy. It should be noted that the diverse ranker component is decoupled from the previously mentioned intent matcher. Existing diverse ranking models, such as those leveraging global document interactions, can be directly applied to implement the diverse ranker, ensuring flexibility in methodology.

D. Training and Optimization

In this section, we describe how to optimize IMEDiv. As mentioned above, in our proposed framework, the intent matcher and diverse ranker are decoupled and can be trained separately. Specifically, for the intent-matching component, we optimized the model with InfoNCE loss:

$$\mathcal{L}_{\text{intent}}(q, d^+, D_N) = -\log \frac{\exp(\text{Intent}(q, \mathcal{I}, d^+))}{\exp(\text{Intent}(q, \mathcal{I}, d^+)) + \sum_{d_i^- \in D_N} \exp(\text{Intent}(q, \mathcal{I}, d_i^-))}, \tag{11}$$

where d^+ represents a positive document and D_N denotes a set of negative documents corresponding to query q. Both d^+ and D_N are sampled from the initial ranking list of q. For the task of intent matching, the positive document d^+ covers more user intents than every negative document $d_i^- \in D_N$. Every document that covers user intents can be the positive document d^+, and the negative set D_N is sampled from all candidate documents in the initial ranking list that cover fewer or no intents compared to d^+. The optimization strategy of intent-matching component aims to allow the model to judge if a document is more likely to cover more user intents with the given query and subtopics.

For the diverse ranker, we use a list-pairwise sampling approach to generate enough training samples in limited datasets. This approach returns a series of positive-negative ranking pairs with common contexts. Consider a pair of samples (r_1, r_2), which can be denoted as $r_1 = [C, d_1]$ and $r_2 = [C, d_2]$, where C is the common context, d_1 is the positive document, and d_2 is the negative document. The context C is generated with different lengths from both an ideal ranking sequence and a randomly sampled sequence. When d_1 or d_2 is appended to the context C, the evaluation metric(i.e., α-nDCG) of positive ranking $\mathcal{M}(r_1)$ should be higher than the negative ranking $\mathcal{M}(r_2)$. The loss function can be defined as a binary classification log-loss function:

$$\mathcal{L} = \sum_{q \in Q} \sum_{s \in S_q} |\Delta\mathcal{M}| \big[y_s \log(P(r_1, r_2)) + (1 - y_s) \log(1 - P(r_1, r_2)) \big]. \tag{12}$$

Here $|\Delta\mathcal{M}| = |\mathcal{M}(r_1) - \mathcal{M}(r_2)|$, and $P(r_1, r_2) = \sigma(s_{r_1} - s_{r_2})$, where $\sigma(\cdot)$ is the sigmoid function. The list-pairwise loss function can be seen as a simulation of a greedy document selection process. With a given context document sequence, the training process with positive-negative sample document pairs optimizes the model for selecting the document, which can lead to a higher evaluation metric value, indicating that the model should select the document that is more beneficial to the diversity of the document sequence.

In conclusion, while previous works focused on enhancing the diverse ranking component with advanced neural networks, IMEDiv pays more attention to improving the intent matching component with an LLM-based ranking model. For the diverse ranking component, IMEDiv only uses an unsupervised max-pooling structure to re-rank the documents. In the following section, our experiments will show that even using a naive ranking model, IMEDiv can still perform well with rich subtopic signals. Besides, IMEDiv is flexible, and it can be integrated into other diversification frameworks, such as DSSA [3] or CL4DIV [13], working as a single intent matching component to achieve better performance.

IV. Experiments

A. Experiment Settings

We use two public datasets, including the **Web Track dataset from TREC 2009-2012** (198 queries) and the **ClueWeb09 dataset** [32]. We used Google query suggestions as first-level subtopics [33]. For the intent-matching component, we used the original query and document content, with documents parsed by Boilerpipe and truncated to the first 320 tokens to reduce computation. The diverse-ranking component

used 18 traditional IR features (e.g., BM25, TF-IDF) provided in [3]. We used the official Web Track diversity metrics: **α-nDCG@20**, **ERR-IA@20**, and **NRBP@20** to evaluate these methods. The top-50 documents from initial rankings were used as input, and all metrics were computed on the top 20 results of the diversified lists. Following prior work, we report two-tailed paired t-tests with $p < 0.05$ to compare methods.

In the training phase, we use 5-fold cross-validation on the datasets. In each fold, we use 4 folds for training and 1 fold for testing, and tune hyperparameters with the widely used metric α-nDCG@20. We compare IMEDiv with all the baselines, including non-diversified approaches and previous implicit and explicit supervised models. For the intent-matching component, we use FlagEmbedding[1] to fine-tune the pretrained model of Gemma-7b-it [34] and apply LoRA [35] with all the linear layers, where the learning rate is set to 2e-4 with rank and alpha set to 16 and 32. While for the diverse ranking component, the learning rate is 1e-3. All the models are trained on a single NVIDIA A800 GPU to ensure consistency.

To address the effectiveness of intent-matching, we conduct two experimental settings of IMEDiv: **IMEDiv (Intent)** denotes that we only use the LLM-based intent-matching component working as a pointwise reranker, returning the intent-enhanced ranking lists $\mathcal{R}_{\text{intent}}$ as the final ranking result with no diverse ranking. **IMEDiv (Intent+Div)** means that the $\mathcal{R}_{\text{intent}}$ is further reranked with the diverse ranking component. Besides, we also conduct another setting named **IMEDiv (Self-Train)**, indicating that the LLM-based reranker is evaluated on training queries. The performance of IMEDiv (Self-Train) serves as an *approximate* upper bound for pointwise intent matching under the same data and model setting, and is reported only for reference. The baseline models are:

- **Lemur**. Following previous works [3], [5], we use the search results produced by the language model and retrieved by Lemur [33] as the non-diversified baseline.
- **xQuAD** [14], **PM2** [15], **HxQuAD** and **HPM2** [33]. These unsupervised explicit approaches compute ranking scores via a linear combination of relevance and diversity, where the methods HxQuAD and HPM2 specifically incorporate hierarchical subtopic structures.
- **R-LTR** [36], **PAMM** [37], **PAMM-NTN** [38] and **Graph4Div** [12][2]. Following PAMM [37], we use the metric of α-nDCG@20 to tune the parameters and use 100-dimensional vectors generated by LDA [39] as the distributed representations of the documents, where the neural tensor network (NTN) is used with R-LTR and PAMM, denoted as **R-LTR-NTN** and **PAMM-NTN**.
- **DSSA** [3]. DSSA diversifies search results by modeling subtopic distributions using LSTM with doc2vec. We use the same settings as reported in the original paper[3].
- **DVGAN** [4]. DVGAN utilizes DSSA as the generator and R-LTR as the discriminator to enhance the search, where the settings used are the same as DSSA.
- **DESA** [5] and **CL4DIV** [13]. DESA uses an encoder-decoder to model document interactions, while CL4DIV employs contrastive learning at the subtopic, with both approaches using doc2vec embeddings. We set the hyperparameters $L_{\text{enc}} = 2$, $L_{\text{dec}} = 1$.

[1]https://github.com/FlagOpen/FlagEmbedding
[2]https://github.com/su-zhan/Graph4DIV
[3]https://github.com/jzbjyb/dssa

B. Overall Results and Discussion

TABLE I: Performance of all approaches (all metrics computed at top-20). Best results are in bold. * indicates statistical significance (p-value<0.05). IMEDiv (Self-Train) represents an approximate upper bound (for reference).

Method	ERR-IA@20	α-nDCG@20	NRBP@20
Lemur	.271	.369	.232
XQuAD	.317	.413	.284
PM2	.306	.411	.267
HxQuAD	.326	.421	.294
HPM2	.317	.420	.279
Graph4Div	.370	.468	.338
CL4Div	.393	.486	.364
R-LTR	.303	.403	.267
PAMM	.309	.411	.271
R-LTR-NTN	.312	.415	.272
PAMM-NTN	.311	.417	.272
DSSA (doc2vec)	.350	.452	.318
DVGAN	.367	.465	.334
DESA	.363	.464	.332
IMEDiv (Intent)	**.376***	**.467***	**.350**
IMEDiv (Intent+Div)	**.375***	**.476***	**.342**
IMEDiv (Self-Train)	**.378**	**.481**	**.348**

Table I reports the overall results. IMEDiv (Intent) and IMEDiv (Intent+Div) outperform most baselines across implicit, explicit, and ensemble categories, and IMEDiv (Intent+Div) is competitive with a recent advanced approach, CL4DIV. These results support our central claim: *improving intent matching is a highly effective and often overlooked lever for better diversification*. Compared with DSSA, which emphasizes a stronger diverse ranker, IMEDiv improves the *upstream* intent signal. Notably, DSSA uses an RNN-based diverse ranker, whereas IMEDiv adopts a simple max-pooling selector; nevertheless, IMEDiv achieves significant gains. This suggests that, under noisy intent signals, designing a better intent matcher can be more beneficial than increasing the complexity of document-interaction modeling.

IMEDiv (Intent) is a pointwise setting that focuses solely on intent coverage: given a query and its subtopics, a document receives a higher score if it is more likely to cover more user intents, regardless of novelty. Despite ignoring novelty, IMEDiv (Intent) still outperforms many listwise diversification models, highlighting the strength of the learned intent signal. Finally, IMEDiv (Self-Train) provides an approximate upper bound when the intent matcher is evaluated on training queries; its strong performance further indicates the potential headroom of intent-centric pointwise reranking for diversification.

C. Effect of Different Pretrained Language Models

We further investigate the effect of different model settings in IMEDiv. As we described in Section III-B, we use decoder-based large language models (LLMs) for instruction tuning to judge if a document covers at least one intent. Here, we investigate the impact of different LLMs for the intent matcher. We use PyTorch [40] and Transformers[4] to implement our framework. The following models are tested in our work: the Gemma [34] models of "gemma-7b-it"[5] and "gemma-2b-it"[6] and the Qwen2.5 [41] models of "qwen-2.5-7b-instruct"[7]. Besides, we also involve the LLM-based reranker of "bge-reranker-v2-gemma"[8] implemented by BAAI [30]. All the model checkpoints are downloaded from Huggingface and ModelScope. These results indicate that the checkpoint of "gemma-7b-it" has achieved the best performance; the LLMs with fewer parameters than 7B may lead to decreased performance. This may be because smaller language models lack good language understanding capabilities and generalization ability. A possible solution is to construct supervised samples to fine-tune the model, which we leave as future work.

TABLE II: Effects of different LLM checkpoints for IMEDiv.

Checkpoints	ERR-IA@20	α-nDCG@20	NRBP@20
qwen2.5-7b-instruct	.347	.446	.315
gemma-7b-it	.376	.467	.350
gemma-2b-it	.342	.439	.310
bge-reranker-v2-gemma	.343	.442	.312

D. Influence of Intent Matching in Diversification

We argue that intent-matching is of great importance in the diversification task, and even an effective diverse ranking component cannot perform well with noisy intent signals. Inspired by the previous work of DESA [5], we analyze effectiveness at *top ranks*. Specifically, we report α-nDCG computed at top-1, top-3, and top-5 to characterize position-aware utility. We compare IMEDiv (Intent) with DSSA, and Table III summarizes the results.

These results indicate that IMEDiv's gains largely come from a stronger intent-matching component. Diversification is most valuable when multiple intents are covered early in the ranked list. Ideally, every document in the initial candidate list should cover at least one intent, but this is rarely true in practice. A list that is "diversified" yet irrelevant provides no utility and receives near-zero diversification metrics. In contrast, a relevant document satisfies at least one intent, and satisfying redundant intents is still preferable to satisfying none; therefore, improving intent matching can directly improve measured diversity.

We further show the top-5 ranking result of query #198 as a case study. For the metric α-nDCG@5 of this query, the result of DSSA is 0, and the result of IMEDiv is 0.275. The results of ranking lists satisfying different user intents are shown in Table IV. From this table, we can see that d_1, d_3, and d_5 of IMEDiv satisfy the intent #1. Although IMEDiv fails to re-rank documents with other intents into the top-5 ranking list, it still outperforms DSSA since DSSA fails to satisfy any of the intents. In the view of the α-nDCG metric, if a ranking list covers some redundant intents, it will be punished and get a lower metric, but if a ranking list covers no intents, the metric will be 0. This analysis is identical to the leading relevance metric nDCG@5 of IMEDiv. These results validate our assumption that enhancing the intent matching component in a diversification framework is of great importance. Our IMEDiv is powerful in intent-matching, and the improvement of intent coverage can effectively enhance the performance of diversification, as more documents satisfy user intents.

TABLE III: Metrics improvement at former ranking positions.

Model	α-nDCG@1	α-nDCG@3	α-nDCG@5
DSSA	.363	.386	.406
IMEDiv	**.400**	**.408**	**.422**

TABLE IV: Case study for IMEDiv (Intent) and DSSA. "Intent satisfied IMEDiv" are user intents satisfied by IMEDiv's documents; "Intent satisfied DSSA" for DSSA's documents.

Document	Intent Satisfied DSSA	Intent Satisfied IMEDiv
d_1	-	#1
d_2	-	-
d_3	-	#1
d_4	-	-
d_5	-	#1

V. Conclusion

This study proposes **IMEDiv**, a supervised diversification framework that improves search result diversity via an enhanced intent matching module. The approach employs an LLM as a pointwise reranker to evaluate document-level intent coverage, decoupling intent matching from later ranking stages to simplify optimization. Experiments show that IMEDiv outperforms strong supervised baselines and remains competitive with recent ensemble methods. It significantly strengthens intent coverage at top ranks, which is a key driver of diversity gains, and its modular design allows "plug-and-play" integration into existing systems. A case study further confirms that relevance matching is essential, as irrelevant documents reduce utility while relevant ones inherently address user intents. Although the current implementation uses greedy selection for diversification, future work may explore joint optimization of document interactions, adaptive fusion for robustness, and multi-objective balancing of relevance, diversity, and novelty.

Acknowledgment

This work was partly supported by the Beijing Natural Science Foundation (No. 4254079), and the National Natural Science Foundation of China Grant (No. 62502028).

[4]https://github.com/huggingface/transformers
[5]https://huggingface.co/google/gemma-7b-it
[6]https://huggingface.co/google/gemma-2b-it
[7]https://www.modelscope.cn/models/Qwen/Qwen2.5-7B-Instruct
[8]https://www.modelscope.cn/models/BAAI/bge-reranker-v2-gemma

REFERENCES

[1] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen, "A large-scale evaluation and analysis of personalized search strategies," in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 581–590.

[2] Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, et al., "Identifying ambiguous queries in web search," in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 1169–1170.

[3] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, et al., "Learning to diversify search results via subtopic attention," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 545–554.

[4] Jiongnan Liu, Zhicheng Dou, Xiaojie Wang, et al., "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*, 2020, pp. 479–488.

[5] Xubo Qin, Zhicheng Dou, and Ji-Rong Wen, "Diversifying search results using self-attention network," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1265–1274.

[6] Quoc Le and Tomas Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 2014, pp. 1188–1196.

[7] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, and et al., "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 2013, pp. 2333–2338.

[8] Liang Pang, Qingyao Ai, and Jun Xu, "Beyond probability ranking principle: Modeling the dependencies among documents," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 1137–1140.

[9] Jaime Carbonell and Jade Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st International ACM SIGIR Conference*, 1998, pp. 335–336.

[10] Zhichao Huang, Xutao Li, Yunming Ye, et al., "Mr-gcn: multi-relational graph convolutional networks based on generalized tensor product," in *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, 2020, pp. 1258–1264.

[11] Zhichao Huang, Xutao Li, Yunming Ye, et al., "Multi-view knowledge graph fusion via knowledge-aware attentional graph neural network," *Applied Intelligence*, vol. 53, no. 4, pp. 3652–3671, 2023.

[12] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen, "Modeling intent graph for search result diversification," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 736–746.

[13] Zhirui Deng, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen, "Cl4div: A contrastive learning framework for search result diversification," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 171–180.

[14] Rodrygo LT Santos et al., "Exploiting query reformulations for web search result diversification," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 881–890.

[15] Van Dang and W Bruce Croft, "Diversity by proportionality: an election-based approach to search result diversification," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, pp. 65–74.

[16] Mehmet Akcay, Ismail Sengor Altingovde, Craig Macdonald, and Iadh Ounis, "On the additivity and weak baselines for search result diversification research," in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 2017, pp. 109–116.

[17] Zhuyun Dai and Jamie Callan, "Deeper text understanding for ir with contextual neural language modeling," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 985–988.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[19] Josh Achiam, Steven Adler, and et al., "Gpt-4 technical report," 2024.

[20] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al., "A survey of large language models," 2023.

[21] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin, "Zero-shot listwise document reranking with a large language model," 2023.

[22] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, and et al., "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*, 2023, pp. 14918–14937.

[23] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, and et al., "Large language models are effective text rankers with pairwise ranking prompting," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 1504–1518.

[24] Xubo Qin, Jun Bai, Jiaqi Li, Zixia Jia, and Zilong Zheng, "Reinforced query reasoners for reasoning-intensive retrieval tasks," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 21261–21274.

[25] Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz, "A survey on rag with llms," *Procedia computer science*, vol. 246, pp. 3781–3790, 2024.

[26] Chunyang Li, Yanping Sun, Zhichao Huang, and Jinjin Guo, "Termrag: Data resource summarization via term retrieval augmented generation," in *International Conference on Intelligent Computing*. Springer, 2025, pp. 225–237.

[27] Yingxu Wang, Shiqi Fan, Mengzhu Wang, et al., "Dynamically adaptive reasoning via llm-guided mcts for efficient and context-aware kgqa," *arXiv preprint arXiv:2508.00719*, 2025.

[28] Wei Ju, Siyu Yi, Yifan Wang, Zhiping Xiao, et al., "A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges," *arXiv preprint arXiv:2403.04468*, 2024.

[29] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin, "Fine-tuning llama for multi-stage text retrieval," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2421–2425.

[30] Jianlyu Chen, Shitao Xiao, Peitian Zhang, et al., "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 2318–2335.

[31] Sadegh Kharazmi, Falk Scholer, David Vallet, and Mark Sanderson, "Examining additivity and weak baselines," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 4, pp. 1–18, 2016.

[32] Charles L Clarke, Nick Craswell, and Ian Soboroff, "Overview of the trec 2009 web track," Tech. Rep., DTIC Document, 2009.

[33] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen, "Search result diversification based on hierarchical intents," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 63–72.

[34] Gemma Team, Thomas Mesnard, Cassidy Hardin, and et al., "Gemma: Open models based on gemini research and technology," 2024.

[35] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, et al., "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[36] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu, "Learning for search result diversification," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014, pp. 293–302.

[37] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng, "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*, 2015, pp. 113–122.

[38] Long Xia, Jun Xu, Yanyan Lan, et al., "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*, 2016, pp. 395–404.

[39] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[40] Adam Paszke, Sam Gross, Francisco Massa, and et al., "Pytorch: an imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 8026–8037.

[41] An Yang, Baosong Yang, Beichen Zhang, and et al., "Qwen2.5 technical report," 2025.