A Novel Matching Paradigm: Unified Generative and Discriminative Large Language Model with Plug-and-Play Fine-Tuning

Anonymous ACL submission

Abstract

Matching paradigm plays a crucial role in large-scale information retrieval and has been widely deployed in industrial search engines. Limited by the feature interaction ability and discriminative architecture, the traditional two-tower and single-tower matching paradigms can no longer satisfy the demands for the performance and interpretability of matching paradigms in the era of LLMs. Existing approaches attempt to utilize LLMs merely as feature extractors, which falls short of fully leveraging the capabilities of LLMs. Therefore, we propose a novel matching paradigm: unified generative and discriminative large language model with plug-and-play fine-tuning (UGD). It integrates the two-tower, single-tower and generative tasks within the same LLM framework through the attention map partition, so as to achieve the deep traction of generative tasks to discriminative tasks and the distillation of single-tower to two-tower discrimination by the plug-and-play multi-task fine-tuning mechanism. To support the training of UGD, we also reconstruct six text matching datasets by appending reason labels based on ERNIE-4.0-Turbo-8K. Extensive experimental results demonstrate that UGD has far superior performance and comparable interpretability. And it has been applied to the industrial search engine, leading to a remarkable enhancement of search experience. Open access upon publication.

1 Introduction

With the explosion of information on the Internet, search and recommendation have become essential ways to meet users' information needs (Kobayashi and Takeda, 2000; Adomavicius and Tuzhilin, 2005). Despite their distinct objectives, search and recommendation can be unified as the "matching" problem from a technical perspective (Garcia-Molina et al., 2011). In industrial search engines and recommendation systems, matching models not only need to obtain high accuracy in recall, but also require efficient retrieval to achieve low latency (Su et al., 2023). Two-tower and single-tower matching have been proven to be effective for industrial large-scale information matching problem (Huang et al., 2020; Yang et al., 2020; Yu et al., 2021; Jang et al., 2023).

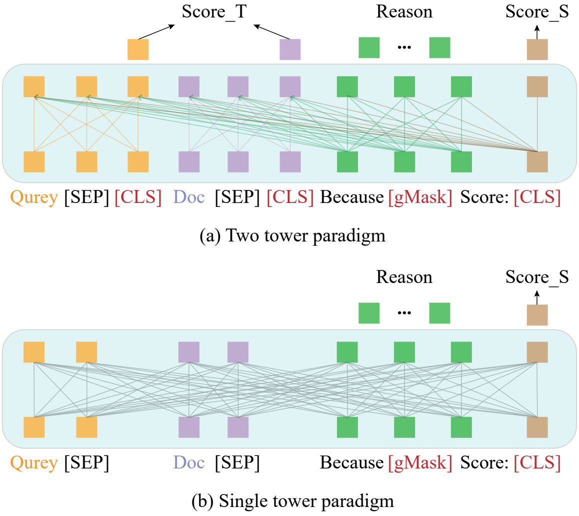

Score_T
Reason
Score_S
Qurey [SEP] [CLS] Doc [SEP] [CLS] Because [gMask] Score: [CLS]
(a) Two tower paradigm
Reason
Score_S
Qurey [SEP] Doc [SEP] Because [gMask] Score: [CLS]
(b) Single tower paradigm

Figure 1: An illustration of unified generative and discriminative matching paradigm.

As we all know, the two-tower model makes use of two encoders to achieve query and document representation respectively, and obtains the relevant score via simple fully connected layer or dot product interaction at the top layer (Reimers, 2019; Huang et al., 2013). However, the fundamental drawback of the traditional two-tower paradigm is manifested in the low matching accuracy due to its limited feature interaction ability. Although the single-tower model interacts with query and document from the bottom layer, it confronts formidable challenges in surmounting the bottleneck of model accuracy due to the limitation of the discriminative architecture (Lu et al., 2022; Devlin, 2018; Liu, 2019). Furthermore, the traditional matching paradigm is a typical black box model, characterized by a lack of interpretability. It implies that users are merely able to obtain the correlation score between the query and the document, yet have no means of discerning the reasons for the correlation discrimination. In numerous fields, such as advertising, e-commerce and search, gaining insights into the reasons for correlation discrimination is of crucial importance. Take search advertising as an example. Advertisers not only need to know the correlation score between user queries and their landing pages but also the underlying reasons. This knowledge empowers them to continuously refine their products and the homepages of their official websites, thereby enhancing their competitiveness and user-friendliness in the market. With the rapid advancements in natural language processing (NLP) technologies, especially with the advent of large language models (LLMs) as elaborated (Li et al., 2023), the traditional matching paradigm needs to augment its abilities and improve its accuracy with the help of LLMs. Moreover, despite some approaches (Ma et al., 2024; Muennighoff et al., 2024) attempt to use generative LLMs to extract features from queries/users and documents/items, it still regards the LLM as an encoder for extracting feature representation, which fails to fully exploit the comprehensive capabilities of the LLM. Therefore, it is particularly important to innovate the traditional matching paradigm based on generative LLMs.

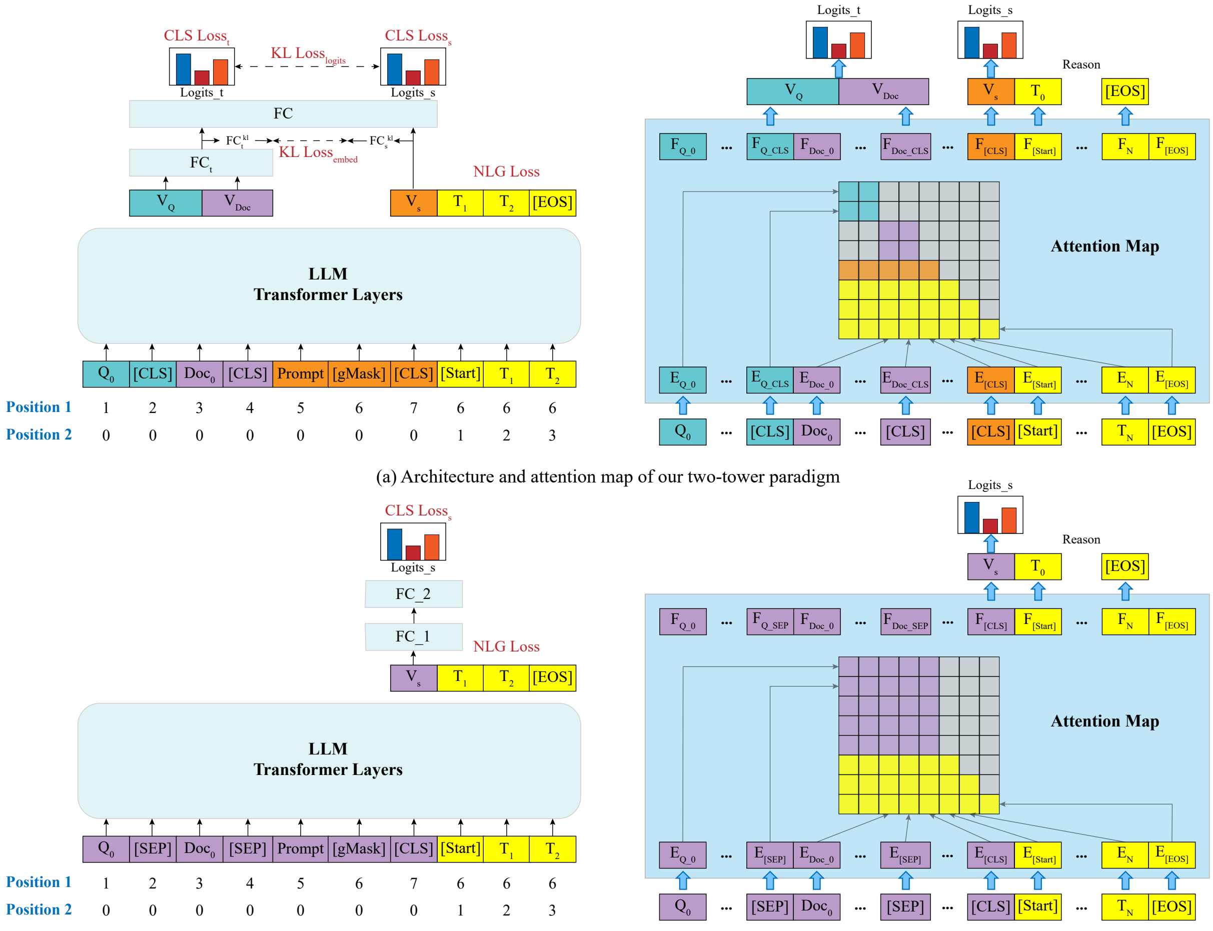
To address the aforementioned challenges, we propose a novel matching paradigm: unified generative and discriminative large language model with plug-and-play fine-tuning (UGD). As shown in Figure 1, it seamlessly integrates the two-tower, single-tower and generative tasks within the same LLM framework through the innovative utilization of attention map partition. Moreover, by means of the plug-and-play multi-task fine-tuning mechanism, it effectively realizes the deep traction of generative tasks to discriminative tasks and the distillation of single-tower to two-tower discrimination. Functionally, our UGD is equipped with generative capabilities. During the inference stage, the generative tasks and discriminative tasks, which are treated as plug-in tasks, can be independently inferred based on the requirements of different scenarios. This independent inference does not impede the inference speed of either the two-tower or single-tower tasks. Technologically, by leveraging the Chain of Thought approach (Wei et al., 2022), the generative task facilitates a more profound interaction between the query and document to achieve better feature representation. Additionally, through the use of the Kullback-Leibler (KL) Loss with detach operation (Wu et al., 2024), it realizes the distillation of single-tower to two-tower discrimination, which significantly enhances the performance of two-tower discrimination. To adapt to the training of UGD matching paradigm, we append the reason labels to the existing text matching datasets through the prompt engineering depending on the top-tier LLMs (Zhang et al., 2019; Achiam et al., 2023). Our contributions are summarized as follows:

- A novel unified generative and discriminative matching paradigm with better matching accuracy and interpretability is proposed, which implemented through attention map partition.
- We develop a plug-and-play multi-task fine-tuning approach, which achieves mutual traction between generative and discriminative tasks, as well as the distillation of single-tower to two-tower discrimination.
- We reconstruct several text matching datasets with discriminative label and generative label based on ERNIE-4.0-Turbo-8K through prompt engineering. Experiments indicate that our UGD paradigm brings a significant improvement in accuracy and interpretability.

2 Related Work

2.1 Traditional Matching Paradigm

From the perspective of the interaction pattern, the traditional matching paradigm can be categorized into two-tower matching and single-tower matching. The two-tower matching is the dominant paradigm in dense retrieval (Huang et al., 2013; Reimers, 2019), which is widely deployed in various applications (Covington et al., 2016; Yi et al., 2019). However, the two-tower matching is confronted with the issue of limited feature interaction ability. To overcome this problem, some researchers proposed introducing SEBlock to enhance feature representation (Wang et al., 2020) and ResNet to make important information more effectively conveyed and fused (Shan et al., 2016). Nevertheless, the improvement is not significant. Other researchers proposed the knowledge distillation approach to better transfer knowledge from the single-tower to the two-tower model (Lin et al.,



(a) Architecture and attention map of our two-tower paradigm

(b) Architecture and attention map of our single-tower paradigm

Figure 2: An overview of the unified generative and discriminative two-tower matching paradigm. In the attention map, the green and purple squares represent the attention mask of all-to-all interaction within query and document respectively. The orange square represents the unidirectional interaction between the single-tower [CLS] token and query and document. The yellow square indicates the attention mask of the generated task. Grey squares are masked out. [CLS] tokens represent the features of the discriminative task and [gMask] token indicates the content to be generated. 2D positional encoding represents inter- and intra-span positions.

2023). However, training single-tower and two-tower models separately for distillation will result in wastage of training resources and inconsistency. In contrast, It has been verified that the single-tower model performs significantly better than the two-tower model due to its all-to-all feature interaction pattern (Kim et al., 2021). Nevertheless, the traditional matching paradigm falls short of meeting the demands of the LLM era.

2.2 General Language Model

With the continuous development of general language model research, it provides technical support to achieve unified generative and discriminative matching paradigm. The general language model aims to achieve the best for all tasks of three main categories including natural language understanding (NLU), unconditional generation and conditional generation, such as UniLM (Dong et al., 2019) and GLM (Du et al., 2021). Inspired by GLM, we have implemented attention map partition to unify the generative and discriminative task into the same LLM. Similarly, both single-tower and two-tower discrimination have also been unified into the same LLM.

3 Method

In this section, we introduce our unified generative and discriminative matching paradigm with plug-and-play fine-tuning. An overview of UGD matching paradigm is presented in Figure 2, which consists of three components: (1) Customized prompt and attention map partition (section 3.1); (2) Model architecture (section 3.2); (3) Plug-and-play multi-task fine-tuning and inference (section 3.3).

3.1 Customized Prompt and Attention Map Partition

Distinguished from traditional matching paradigms, our method employs customized CoT-prompt and partitioned attention map as inputs.

First of all, as shown in Figure 2(a), we take query-document pairs and correlation discrimination reasons as inputs for generative LLM through CoT prompt engineering. Specifically, the input is structured as follows: "query + [CLS] + document + [CLS] + because [gMask], the correlation score is [CLS] + [Start] + reason + [EOS]". Here, query and document are the content to be discriminated, and the first two corresponding special tokens ([CLS]) are employed to denote their feature representations. To ensure the independence of the query and document of the two-tower model, as shown on the right-hand side of Figure 2(a), we adopt an attention map to achieve that the query and document can only interact with themselves. Besides, in order to enhance the interpretability and logicality of the two-tower model, a discriminant reason generative task is added on the basis of the correlation discrimination task, where a special token ([gMask]) represents the reason to be generated. "[Start] + reason + [EOS]" represents the label of the generative task and its attention map is a typical lower triangular matrix. Finally, to realize the knowledge distillation of single-tower over two-tower discrimination, a special token ([CLS]) is added after the reason occupying special token ([gMask]) as the feature representation for the single-tower discriminative task. It is worth noting that the special token ([CLS]) here can access the query, document, and reason occupying token mentioned earlier. Thus, we have successfully achieved the two-tower UGD paradigm through customized prompts and attention map partition. As depicted in Figure 2(b), single-tower UGD paradigm has only one special token ([CLS]) that represents the feature representation of the single-tower discriminative task. It is mutually visible with the previous query, document, and reason occupying token.

3.2 Model Architecture

The model architecture of our UGD matching paradigm, as shown in Figure 2, consists of LLM backbones and multi-task head networks.

LLM Backbones. In this study, we utilize the GLM paradigm as LLM backbone (Du et al., 2021). To ensure that the length and position of the reasons represented by the special token ([gMask]) are not restricted, the 2D positional encoding is employed. As depicted on the left of Figure 2(a), specifically, each token is encoded with two positional ids. The first positional id represents the position in the context. For the masked span, it is the position of the corresponding [gMask] token. The second positional id represents the intra-span position (range from 1 to the length of the masked span). Certainly, our method is also applicable to most autoregressive LLMs with their positional encoding, including the series of Llama (Touvron et al., 2023), Qwen (Yang et al., 2024) and Ernie (Zhang et al., 2019), etc. As a result, encoder outputs $F \in L * D$ (L represents the maximum length of the input sequence and D denotes the embedding dimension) can be obtained through transformer layers of LLM.

Multi-Task Head Networks. The architecture of our two-tower Unified Generative and Discriminative (UGD) model is illustrated on the left side of Figure 2(a). Evidently, it consists of three head networks, which are specifically designed for two-tower discriminative task, single-tower discriminative task and generative task. As elaborated in section 3.1, we extract the vectors corresponding to the first two special tokens ([CLS]) in the encoder outputs F as the query and document feature representations (namely V_Q and V_{Doc}), respectively. After concatenating V_Q and V_{Doc}, F_t is obtained by dimension reduction through FC_t. F_t is processed by FC and FC_t^{kl} to obtain $Logits_t$ and $\hat{F}_t$, respectively. Analogously, the feature representation of the single-tower discriminative task is denoted as V_s. V_s is processed by FC and FC_s^{kl} to obtain $Logits_s$ and $\hat{F}_s$, respectively. Among them, $Logits_s$ and $\hat{F}_s$ of the single tower are used to perform KL constraints on $Logits_t$ and $\hat{F}_t$ of the two tower for knowledge distillation. Besides, for the generative task, the LMHead layer maps encoder outputs F from the embedding dimension D to the vocabulary size C_{vocab}.

3.3 Plug-and-Play Multi-Task Fine-Tuning and Inference

In this section, during the fine-tuning stage, we design multiple tasks as plugins to fine-tune the objective function, which is composed of multiple loss functions. These tasks collaborate and optimize with one another to improve matching accuracy while ensuring consistency. During the inference stage, the necessary task plugins can be selected for the requirements of different business scenarios, thus achieving a plug-and-play effect. It not only improves the adaptability to various practical applications but also streamlines the deployment process, enabling more efficient utilization of the capabilities in different business scenarios.

As for the two-tower UGD paradigm, the objective function is composed of five loss functions in Figure 2(a). Firstly, $CrossEntropyLoss$ is for the two-tower and single-tower discriminative tasks and the reason generative task.

$$\mathcal{L}_{cls}^{t} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\sum_{i=1}^{C} y_i \log(p_i^t)\right] \quad (1)$$

$$\mathcal{L}_{cls}^{s} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\sum_{i=1}^{C} y_i \log(p_i^s)\right] \quad (2)$$

$$\mathcal{L}_{gen} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\frac{1}{T}\sum_{t=1}^{T} \log q(y_t|y_{<t})\right] \quad (3)$$

$KullbackLeiblerLoss$ aims to achieve distillation of the single-tower model to the two-tower model. The loss functions are as follows:

$$\begin{aligned}\mathcal{L}_{kl} &= D(P\|Q.detach)\\ &= \sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right)\end{aligned} \quad (4)$$

$$\begin{aligned}\mathcal{L}_{kl}^{emb} &= D(T\|S.detach)\\ &= \sum_{x\in X} T(x)\log\left(\frac{T(x)}{S(x)}\right)\end{aligned} \quad (5)$$

where $\mathcal{L}_{cls}^{t}$ and $\mathcal{L}_{cls}^{s}$ represent the loss functions of the two-tower and single-tower discrimination, respectively. $y_i \in (y_1, ..., y_C)$ represents the label of the discriminative task. $p_i^t \in Logits_t = (p_1^t, ..., p_C^t)$. C is the number of categories for the discriminative task. Similarly, $p_i^s \in Logits_s = (p_1^s, ..., p_C^s)$. L_{gen} represents the loss function of the generative task. Besides, $\mathcal{L}_{kl}$ and $\mathcal{L}_{kl}^{emb}$ are the KL Loss function for the distillation of single-tower to two-tower discrimination, respectively for $logits$ and $embeddings$. $P = Logits_t$ and $Q = Logits_s$ aim at the distillation of $logits$. Correspondingly, $T = \hat{F}_t$ and $S = \hat{F}_s$ aim at the distillation of $embeddings$. Therefore, the objective function of our two-tower UGD paradigm is as follows:

$$\mathcal{L}_t = \alpha\mathcal{L}_{cls}^{t} + \beta\mathcal{L}_{cls}^{s} + \gamma\mathcal{L}_{gen} + \lambda\mathcal{L}_{kl} + \mu\mathcal{L}_{kl}^{emb} \quad (6)$$

where $\{\alpha, \beta, \gamma, \lambda, \mu\}$ represents the weight coefficients of each loss function. In our experiment, the parameters were set as $\{\alpha = 1, \beta = 1, \gamma = 1, \lambda = 10, \mu = 10\}$. Similarly, the objective function of the single-tower UGD paradigm is composed of $\mathcal{L}_{cls}^{s}$ and $\mathcal{L}_{gen}$ as follows:

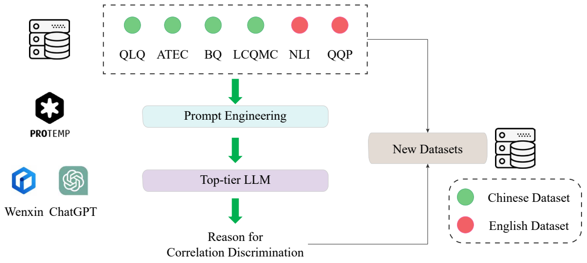


Figure 3: A flowchart for generating reason labels based on the top-tier LLMs.

$$\mathcal{L}_s = \beta\mathcal{L}_{cls}^{s} + \gamma\mathcal{L}_{gen} \quad (7)$$

where $\{\beta = 1, \gamma = 1\}$.

It is worth noting that in the Kullback-Leibler Loss, we independently conduct the $detach$ (Van Den Oord et al., 2017) operation on $Logits_s$ (denoted as Q) and $\hat{F}_s$ (denoted as S). The mechanism of Kullback Leibler Loss with $detach$ is to optimize only the two-tower module so that they can approximate the performance of the single-tower module, as opposed to engaging in bidirectional optimization. Besides, applying constraints to both Q and its preceding feature vector S leads to enhanced effectiveness in knowledge distillation. This targeted optimization approach allows for more precise control over the learning process, enabling the two-tower architecture to better mimic the characteristics of the single-tower architecture, which lead to improved overall model performance and more effective knowledge transfer.

4 Experiments

4.1 Datasets and their Reconstruction

In this section, existing datasets and their reconstruction processes are introduced to adapt to UGD matching paradigm.

Existing datasets. Four Chinese datasets and two English datasets are collected for subsequent experiments. Chinese datasets include: (1) **QLQ dataset** is created by ourself and utilized to to assess the quality of queries and landing pages in the search advertising domain. It provides user with queries and landing page descriptions, along with their evaluation quality levels (scaled from 0 to 3) and corresponding reasons. The training set has $2311,326$ samples, while the testing set contains

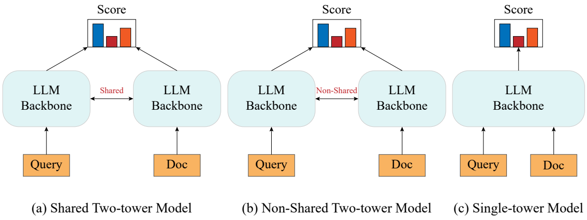
36,885 samples. (2) **ATEC dataset** is a financial semantic similarity dataset. It provides pairs of financial questions and their semantic relevance labels (either 0 or 1). The training set has 82,477 samples and the testing set has 20,000 samples. (3) **BQCorpus dataset** is a Chinese corpus for sentence semantic equivalence identification in the field of bank finance created by (Chen et al., 2018). The BQ corpus contains 120,000 question pairs with the semantic relevance label (0 or 1) from 1-year online bank custom service logs. Among these, 100,000 samples are in the training set and 20,000 in the testing set. (4) **LCQMC dataset** is a large-scale Chinese question matching corpus constructed by (Liu et al., 2018). It consists of a training set of 238,766 and a testing set of 21,302 with the semantic relevance label (0 or 1).

English datasets include: (1) **NLI dataset** is a concatenation of the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) datasets, which covers a wide range of subjects, including novels, government documents, academic papers. The semantic relationship between two sentences is marked as {"0": "entailment", "1": "neutral", "2", "contradiction"}. The training set and testing set contain 941,445 and 39,307 samples. (2) **QQP dataset** (Quora Question Pairs) is a collection of question pairs from the community question-answering website Quora. The task is to determine whether a pair of questions are semantically equivalent (0 or 1). It consists of a training set of 297,708 and a testing set of 32,965.

Dataset reconstruction based on prompt engineering. With the exception of the QLQ dataset that we have provided, the other publicly available datasets only contain category labels and lack the reason label required for the generative task. The remarkable progress of general LLMs provides an extremely convenient solution for data annotation. As shown in Figure 3, we illustrate a flowchart for generating reason labels based on the top-tier LLM (ERNIE-4.0-Turbo-8K). In detail, we extract sentence pairs and their category labels from the original dataset, and employ the prompt engineering technique to construct customized prompts for generating reasons based on top-tier LLMs. The reconstructed datasets are elaborated in the Appendix A and will be made accessible to the public.

4.2 Performance Analysis

To evaluate the results, our proposed UGD matching paradigm is compared with the traditional matching paradigm. For simplicity, we abbreviate the "Shared Two-tower Model" as "Shared-TTM", the "Non-Shared Two-tower Model" as "TTM" and the "Single-tower Model" as "STM" for the subsequent contents, tables and figures.



(a) Shared Two-tower Model (b) Non-Shared Two-tower Model (c) Single-tower Model

Figure 4: The architecture of baseline models.

Baseline models. As presented in Figure 4, we introduce the traditional matching paradigm architecture as baseline models, including Shared-TTM, TTM and STM. For Shared-TTM, the query side and document side share an LLM backbone as the feature extraction module. Conversely, the TTM is equipped with two LLM backbones, and their parameters are not shared. For Shared-TTM and TTM, the query and document are first embedded and then interact, with the interaction occurring only at the top level. In contrast, STM initiates the interaction between query and document from the bottom based on a single LLM backbone.

In our experiments, we choose Ernie 3.0 Zeus 1.5B as the LLM backbone. We deploy full-parameter fine-tuning with the $5e-6$ learning rate, which is implemented on 8 NVIDIA A800 GPUs.

Comparative Experiments. A comparison of the performance of UGD matching paradigm and baselines is presented in Tables 1 and 2. As shown in Tables 1 and 2, across most Chinese and English datasets, our UGD matching paradigm has made significant improvements in ACC and AUC indicators compared with the traditional matching paradigm. It demonstrates that our UGD matching paradigm can better stimulate the capabilities of generative LLMs, rather than just regarding them as feature extractors.

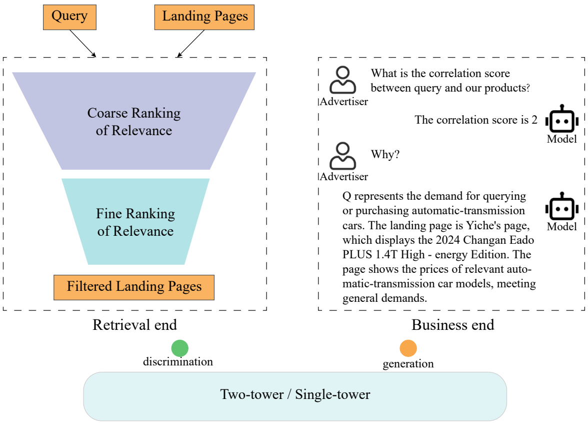
Specially, as for the paradigm of TTM, our UGD TTM integrates the two-tower, single-tower and generative tasks into the same LLM model through the attention map partition, so as to achieve the deep traction of generative tasks to discriminative tasks and the distillation of single-tower discrimination to two-tower discrimination. Therefore, UGD TTM has better performance than Shared TTM and TTM which only take LLM as feature extrac-

Type	Methods	Chinese Datasets (zh) (ACC / AUC)			
		QLQ	BQ	ATEC	LCQMC
TTM	Shared TTM	0.7183 / 0.9141	0.8170 / 0.8892	0.8139 / 0.6940	0.8215 / 0.9182
	TTM	0.7236 / 0.9137	0.7172 / 0.7849	0.8176 / 0.6297	0.7478 / 0.8459
	UGD TTM	**0.7311 / 0.9185**	**0.8448 / 0.9180**	**0.8284 / 0.8288**	**0.8331 / 0.9185**
STM	STM	0.7443 / 0.9343	**0.8614 / 0.9351**	0.8685 / 0.8966	0.8874 / 0.9616
	UGD STM	**0.7525 / 0.9369**	0.8604 / 0.9342	**0.8728 / 0.9022**	**0.8908 / 0.9619**

Table 1: Performance comparisons on Chinese Datasets. LLM backbone is Ernie 3.0 Zeus 1.5b. The best results are in bold.

Type	Methods	English Datasets (en) (ACC / AUC)	
		NLI	QQP
TTM	Shared TTM	0.7646 / 0.9253	0.9223 / 0.9622
	TTM	0.7444 / 0.9130	0.8977 / 0.9397
	UGD TTM	**0.7887 / 0.9388**	**0.9300 / 0.9807**
STM	STM	0.8701 / 0.9744	0.9752 / **0.9905**
	UGD STM	**0.8746 / 0.9786**	**0.9803** / 0.9868

Table 2: Performance comparisons on English Datasets. LLM backbone is Ernie 3.0 Zeus 1.5b. The best results are in bold.



What is the correlation score between query and our products?

The correlation score is 2

Why?

Q represents the demand for querying or purchasing automatic-transmission cars. The landing page is Yiche's page, which displays the 2024 Changan Eado PLUS 1.4T High - energy Edition. The page shows the prices of relevant automatic-transmission car models, meeting general demands.

Figure 5: An illustration of the application of our UGD matching paradigm in search advertising.

tor. Specifically, UGD TTM achieve the ACC of 73.11% and AUC of 91.85% on the search advertisement dataset QLQ, which realize an increase of about 1.0% of ACC compared with shared TTM and TTM. The characteristic of this dataset is that queries are usually short keywords, while documents are long and discrete landing page contents. BQ, ATEC and LCQMC are the corpora of Chinese question pairs, and their two sentence pairs are usually of equal length and short. Better performances are achieved on each of these three datasets with the increases of 2.78%, 1.08%, and 1.16% of ACC than others respectively (ACC / AUC: 84.48% / 91.80%, 82.84% / 82.88%, and 83.31% / 91.85%). In addition, the increases of 2.41% and 0.77% of ACC are achieved compared to baseline models on the English datasets NLI and QQP. Similarly, for the paradigm of STM, benefiting from the promotion of interaction between query and document brought about by generative tasks, our UGD STM achieved an improvement of 0.5% to 1.0% compared to traditional STM on the Chinese and English datasets mentioned above in various fields.

Beyond enhancing performance, our UGD provides a plugin for generating tasks. In our experimental setup, this manifests as the discriminative reason generation. It also endows the model with interpretability, providing reasons of correlation discrimination on the business end. The case study is explained in detail within the Appendix B.

4.3 Online Testing

In this section, we deploy the UGD matching paradigm on the industrial search engine and randomly take a small percentage of traffic as the test group. Figure 5 illustrates the application of the UGD matching paradigm in the search advertising domain. On the left side is the retrieval end, whose responsibility is to filter the landing pages corresponding to a specified query. On the right side is the business end, which is tasked with notifying advertisers of the relevance score and the reasons behind the score between their products and the query. At the retrieval end, the coarse- and fine-rank filtering of candidate landing pages is realized through the discrimination functions of our TTM and STM UGD matching paradigms. At

Model	Multi-Task Loss					Metric
(UGD)	$\mathcal{L}_{cls}^{t}$	$\mathcal{L}_{cls}^{s}$	$\mathcal{L}_{gen}$	$\mathcal{L}_{kl}$	$\mathcal{L}_{kl}^{emb}$	(ACC / AUC)
TTM	✓	✓	✓	w/o	w/o	0.7225 / 0.9140
	✓	✓	✓	✓	✓	**0.7311 / 0.9185**
	✓	✓	✓	✓	×	0.7289 / 0.9185
	✓	✓	✓	×	×	0.7212 / 0.9129
	✓	✓	×	×	×	0.7149 / 0.9074
	✓	×	×	×	×	0.7158 / 0.9087
STM	-	✓	✓	-	-	**0.7525 / 0.9369**
	-	✓	×	-	-	0.7443 / 0.9343

Table 3: Ablation experiments on the QLQ dataset. We conduct ablation experiments for UGD TTM and UGD STM from the loss functions of $\mathcal{L}_{cls}^{t}$, $\mathcal{L}_{cls}^{s}$, $\mathcal{L}_{gen}$, $\mathcal{L}_{kl}$ and $\mathcal{L}_{kl}^{emb}$, as well as their corresponding head networks. w/o represents "KL Loss without detach". The best results are in bold.

the business end, advertisers can gain insights into the reasons for relevance discrimination to improve their products and optimize the landing page design. As a result, a virtuous feedback loop is established. Through continuous improvement based on the provided reasons, the quality of products and landing pages can be steadily enhanced, leading to better user experiences and advertising campaigns.

During the A/B testing period, which typically spans at least one week, we closely monitor the performance of the UGD matching paradigm. Compared with the previously deployed model, the TTM UGD effectively reduces the proportion of 0-score landing pages by **1.87%**. Meanwhile, the STM UGD achieves an even more substantial reduction for the proportion of 0-score landing pages by **3.2%**. Details are in the Appendix C.

4.4 Ablation Study

As shown in Tables 3, we conduct ablation experiments for UGD TTM and STM around the loss functions of $\mathcal{L}_{cls}^{t}$, $\mathcal{L}_{cls}^{s}$, $\mathcal{L}_{gen}$, $\mathcal{L}_{kl}$ and $\mathcal{L}_{kl}^{emb}$, as well as corresponding head networks.

As for UGD TTM matching paradigm, we first conduct ablation experiments on the knowledge distillation of single-tower to two-tower discrimination. The knowledge distillation for both $logits$ and $embedding$ is conducted using KL Loss. When the detach operation is applied in the KL Loss calculation, the performance improves by approximately 1% compared to the case without it (ACC from 72.25% to 73.11%). Because of the detach operation, the model optimizes the model parameters to make the $logits$ and $embedding$ of the two-tower closer to those of the single-tower. When the $\mathcal{L}_{kl}^{emb}$ loss term is removed, ACC drops from 73.11% to 72.89%. Moreover, when the entire knowledge distillation module is removed, ACC drops from 73.11% to 72.12%. They validate the effectiveness of knowledge distillation between single-tower to two-tower discrimination. However, our work is the first to integrate knowledge distillation into one LLM through attention map partition, which saves training resources while ensuring its consistency. Besides, after removing the reason generative task $\mathcal{L}_{gen}$, the ACC of UGD TTM decreases from 72.12% to 71.49% and the ACC of UGD STM decreases from 75.25% to 74.43%, indicating that the generative task has excellent feature traction effect for matching paradigm based on LLMs. Furthermore, Tables 3 also indicates that single-tower discrimination itself does not improve two-tower discrimination. Instead, it must be constrained by KL Loss to achieve knowledge distillation.

5 Conclusion

In this paper, we explore the innovation of traditional matching paradigm in the era of generative LLMs. To improve the performance of traditional matching paradigms and endow them with interpretability, we propose a novel matching paradigm: unified generative and discriminative large language model with plug-and-play fine-tuning. It integrates the two-tower, single-tower and generative tasks into the same LLM model through the attention map partition, so as to achieve the deep traction of generative tasks to discriminative tasks and the distillation of single-tower discrimination to two-tower discrimination by the plug-and-play multi-task fine-tuning mechanism. Moreover, to support the training of UGD, we reconstruct six text matching datasets by appending the reason labels through the prompt engineering based on the top-tier LLM (ERNIE-4.0-Turbo-8K). Extensive experimental results on these datasets show that UGD has far superior performance and comparable interpretability to the traditional matching paradigms. And it has been applied to the industrial search engine and significantly improved the search experience and satisfaction. In the future, we will further explore the application of UGD matching paradigm and its innovations to other tasks, such as multi-modal matching and prompt compression tasks.

6 Limitations

Compared with traditional matching paradigms, our UGD demonstrates superior performance and interpretability. However, this study is not without limitations. In this paper, the LLM backbone adheres to the GLM paradigm. However, causal LLMs require further exploration. Given the constraints of the industrial search engine scenario, the experiments in this paper utilize a 1.5B-parameter LLM backbone. Although, in theory, our proposed paradigm is applicable to larger-scale LLMs, experimental validation is yet to be carried out.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4946–4951.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Hector Garcia-Molina, Georgia Koutrika, and Aditya Parameswaran. 2011. Information seeking: convergence of search, recommendations, and advertising. *Communications of the ACM*, 54(11):121–130.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. 2023. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 980–988.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM computing surveys (CSUR)*, 32(2):144–173.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, et al. 2023. Prod: Progressive distillation for dense retrieval. In *Proceedings of the ACM Web Conference 2023*, pages 3299–3308.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th international conference on computational linguistics*, pages 1952–1962.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 15692–15701.

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

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. 2016. Deep crossing: Web-scale modeling without manually crafted combinatorial features. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 255–262.

Liangcai Su, Fan Yan, Jieming Zhu, Xi Xiao, Haoyi Duan, Zhou Zhao, Zhenhua Dong, and Ruiming Tang. 2023. Beyond two-tower matching: Learning sparse retrievable cross-interactions for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 548–557.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2020. Cold: Towards the next generation of pre-ranking system. *arXiv preprint arXiv:2007.16122*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. 2024. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*.

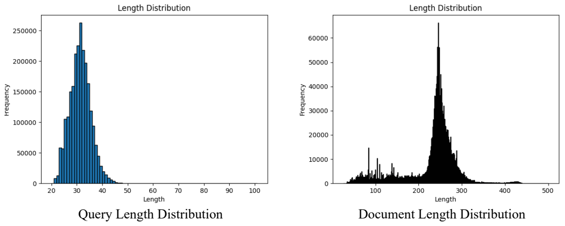
An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H Chi. 2020. Mixed negative sampling for learning two-tower neural networks in recommendations. In *Companion proceedings of the web conference 2020*, pages 441–447.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM conference on recommender systems*, pages 269–277.

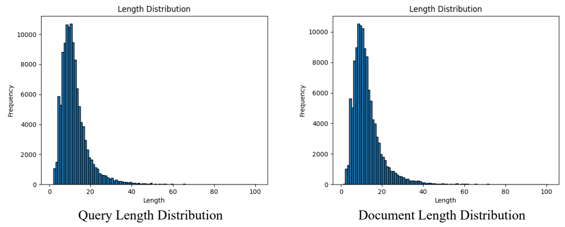
Yantao Yu, Weipeng Wang, Zhoutian Feng, and Daiyue Xue. 2021. A dual augmented two-tower model for online large-scale recommendation. *DLP-KDD*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.



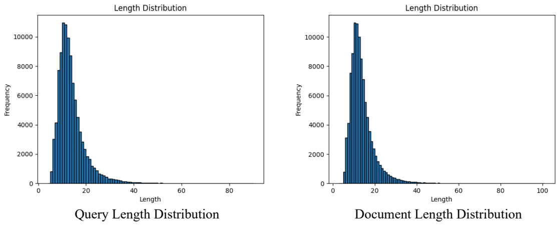
Query Length Distribution Document Length Distribution

Figure 6: An illustration of the distribution of the lengths of queries and documents in Dataset QLQ.



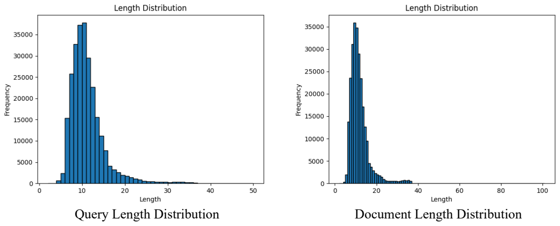
Query Length Distribution Document Length Distribution

Figure 7: An illustration of the distribution of the lengths of queries and documents in Dataset ATEC.



Query Length Distribution Document Length Distribution

Figure 8: An illustration of the distribution of the lengths of queries and documents in Dataset BQ.



Query Length Distribution Document Length Distribution

Figure 9: An illustration of the distribution of the lengths of queries and documents in Dataset LCQMC.

A Appendix for Details of Datasets

Datasets	Train set	Test set	Avg Length			CLS Label
	Size	Size	Query	Document	Reason	
QLQ	2311326	36885	30.87	229.55	146.48	4
BQ	100000	20000	11.63	12.09	74.94	2
ATEC	82477	20000	13.33	13.35	75.23	2
LCQMC	238766	21302	10.68	11.19	69.77	2
NLI	941445	39307	15.81	8.50	58.45	3
QQP	297708	32965	11.13	11.41	55.96	2

Table 4: Detailed information of the dataset.

Tables 4 presents detailed information about the Chinese and English datasets employed in this paper. It includes the sizes of the training and testing sets, along with the average lengths of queries, documents and reasons. Figure 6 to 11 show the length distribution of query and document for different datasets.

B Appendix for Case Study

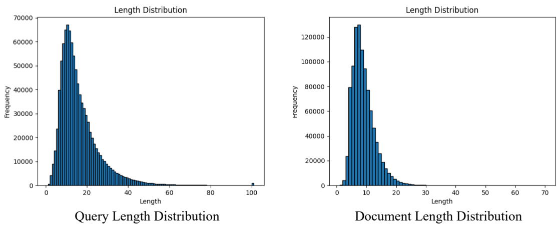
In this section, we conduct a qualitative analysis of our UGD matching paradigm from the perspective of case studies.

As shown in Figure 12, in the field of search advertising, it is necessary to filter out the most relevant landing pages from the candidate set of landing pages based on user queries. Usually, the landing page content is highly discrete and noisy. Compared with traditional matching paradigms, our UGD matching paradigm not only improves discriminative performance but also provides corresponding reasons.

As mentioned in the previous section, ATEC, BQ, and LCQMC are all Chinese question matching datasets. As shown in Figure 13, taking BQ dataset as an example, Through the innovative design of the generation task and knowledge distillation, the model can better understand the meanings of two sentences, thereby providing more logical discrimination results.

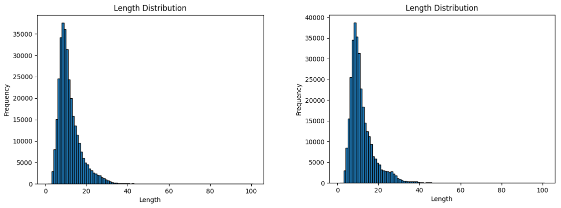
Similarly, we also conduct the case studies on the English dataset NLI in Figure 14. The semantic relationship between two sentences is marked as {"0": "entailment", "1": "neutral", "2", "contradiction"}. This can also prove the above conclusion.

In summary, the innovative integration of generation tasks and knowledge distillation within our UGD matching paradigm has significantly enhanced the model's capacity to comprehend sentence meanings. This improvement has, in turn, enabled the generation of more logically sound discrimination outcomes. The positive results across



Query Length Distribution Document Length Distribution

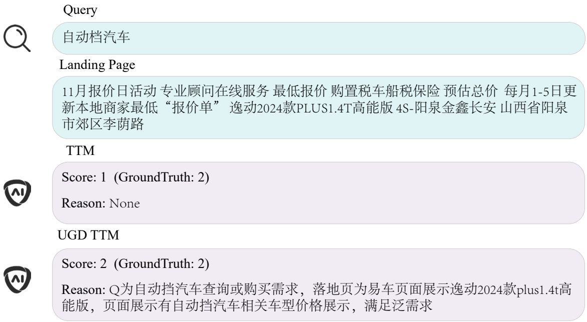
Figure 10: An illustration of the distribution of the lengths of queries and documents in Dataset NLI.



Query Length Distribution

Document Length Distribution

Figure 11: An illustration of the distribution of the lengths of queries and documents in Dataset QQP.



Query

自动档汽车

Landing Page

11月报价日活动 专业顾问在线服务 最低报价 购置税车船税保险 预估总价 每月1-5日更新本地商家最低“报价单” 逸动2024款PLUS1.4T高能版 4S-阳泉金鑫长安 山西省阳泉市郊区李荫路

TTM

Score: 1 (GroundTruth: 2)

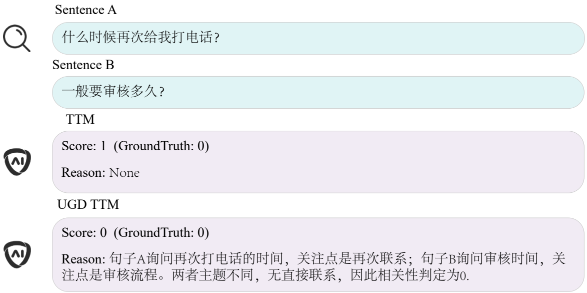
Reason: None

UGD TTM

Score: 2 (GroundTruth: 2)

Reason: Q为自动挡汽车查询或购买需求，落地页为易车页面展示逸动2024款plus1.4t高能版，页面展示有自动挡汽车相关车型价格展示，满足泛需求

Figure 12: The case studies of QLQ dataset.



Sentence A

什么时候再次给我打电话?

Sentence B

一般要审核多久?

TTM

Score: 1 (GroundTruth: 0)

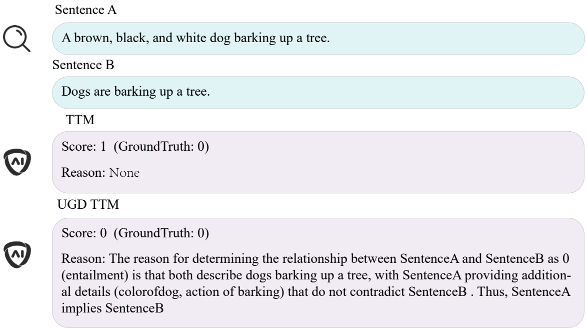
Reason: None

UGD TTM

Score: 0 (GroundTruth: 0)

Reason: 句子A询问再次打电话的时间，关注点是再次联系；句子B询问审核时间，关注点是审核流程。两者主题不同，无直接联系，因此相关性判定为0.

Figure 13: The case studies of BQ dataset.



Sentence A

A brown, black, and white dog barking up a tree.

Sentence B

Dogs are barking up a tree.

TTM

Score: 1 (GroundTruth: 0)

Reason: None

UGD TTM

Score: 0 (GroundTruth: 0)

Reason: The reason for determining the relationship between SentenceA and SentenceB as 0 (entailment) is that both describe dogs barking up a tree, with SentenceA providing additional details (colorofdog, action of barking) that do not contradict SentenceB . Thus, SentenceA implies SentenceB

Figure 14: The case studies of NLI dataset.

a range of experiments validate the superiority and effectiveness of the UGD matching paradigm in diverse matching scenarios, highlighting its potential to outperform traditional approaches and drive progress in related research fields.

C Appendix for Online Testing and Other Applications

During the A/B testing period, we closely monitor the performance of the UGD matching paradigm. Compared with the previously deployed model, the TTM UGD matching paradigm effectively reduces the proportion of 0-score landing pages by **1.87%**. Meanwhile, the STM UGD matching paradigm achieves a **3.9%** increase in filtering accuracy, a **1.6%** increase in filtering volume, a **3.2%** decrease in 0-score landing page display, a **1.36%** decrease in 1-score landing page display, and a **3.09%** increase in 2- and 3-score landing page display.

Furthermore, the proposed approach has broad applicability across diverse fields. In the realm of medical diagnosis assistance, it not only furnishes the confidence levels corresponding to different disease categories but also generates in - depth explanations for symptoms, along with potential diagnostic suggestions. This dual - function approach enriches the diagnostic process, providing medical professionals and patients with more comprehensive information for informed decision-making. Regarding question-answering systems, they are designed to generate accurate answers to user queries while concurrently estimating the probability of correlation between the input questions and the entries stored within the knowledge base. This probability assessment is crucial as it allows for a more refined evaluation of the relevance of the retrieved answers, enhancing the overall quality of the response. In the case of recommendation systems, detailed explanations are provided for the recommended results. These explanations play a pivotal role in helping users understand the underlying rationales behind why specific products or content are being recommended. By offering such transparency, the user experience is significantly improved, leading to increased user trust and potentially higher engagement with the recommended items.

D Appendix for Implementation Details

The proposed UGD was trained through full-parameter fine-tuning. During training, the learning rate was set to $5e-6$, and the batch size was 32. The Adam optimizer with $\beta_1 = 0.9$ and $\beta_1 = 0.999$. The $WarmupSteps$ was set to 4000. It was implemented on 8 NVIDIA A800 GPUs for 5 epochs. The entire training process spanned over 5 days.