general writing skills

when idea is simply and straightforward

- 想法简单，但是可能具体实现有很多困难。
    - In this way, it is able to better learn the similarity relations among query, positive passages and negative passages. Although the idea is appealing, it is not easy to implement due to three major issues. First, it is unclear how to formalize and learn both query-centric and passage-centric similarity relations. Second, it requires arge-scale and highquality training data to incorporate passage-centric similarity relation. However, it is expensive to manually label data. Additionally, there might be a large number of unlabeled positives even in the existing manually labeled datasets (Qu et al., 2020), and it is likely to bring false negatives when sampling hard negatives. Finally, learning passagecentric similarity relation (an auxiliary task) is not directly related to the query-centric similarity relation (a target task). In terms of multi-task viewpoint, multi-task models often perform worse than their single-task counterparts (Alonso and Plank, 2017; McCann et al., 2018; Clark et al., 2019). Hence, it needs a more elaborate design for the training procedure.
    - To this end, in this paper, we propose a novel approach that leverages both query-centric and PAssage-centric sImilarity Relations (called PAIR) for dense passage retrieval. In order to address the aforementioned issues, we have made three important technical contributions. First,
    - 这个范例很好的展示这种情况怎么写。重点说明为什么不好实现或中间具体的困难在哪里，不然的话如果真的那么容易文章就没那么大价值了，谁都可以做吗。避免审稿时，审稿人可能一看可能觉得这个好像也大的创新。

- 方法简单，但是效果好
    - danqi chen: SimCSE
    - This paper presents SimCSE, a simple contrastive learning framework that greatly advances the state-of-the-art sentence embeddings. We first describe an unsupervised approach, which takes an input sentence and predicts itself in a contrastive objective, with only standard dropout used as noise. ***This simple method works surprisingly well, performing on par with previous supervised counterparts.*** We hypothesize that dropout acts as minimal data augmentation and removing it leads to a representation collapse. Then, we draw inspiration from the recent success of learning sentence embeddings from natural language inference (NLI) datasets and incorporate annotated pairs from NLI datasets into contrastive learning by using “entailment” pairs as positives and “contradiction” pairs as hard negatives.
    - 这篇文章想法和实现都简单得不像话，如果功底一般哪怕发现了 SimCSE 中的 tricks 可以提高也发不了顶会，甚至自己都觉得想法太简单，就一个 trick 而已就不些论文了。但是大神就是大神，分析的神出鬼没，感叹一个牛字。做研究基本功很重要。

虽然前人有进展，但是还是值得研究

- Despite the progress made so far, there is still a need for developing a more precise method for peptide–MHC binding prediction to reduce the large number of false positives and thus improve the confidence of the predicted peptide–MHC interactions. In addition, improving the correlations between

predicted and measured binding affinities may help quantify the binding advantage of neoantigens compared to the wild-type version, which can further facilitate vaccine development. Moreover, the prediction results from most of the previous methods

方法有明显不足不能实现，但也有好的方面

- In this paper, DSI is applied to moderate-sized corpora (from 10k to 320k documents), all of which are derived from one challenging retrieval task, and we leave the important question of the scaling DSI to larger corpora to future work.
-

conclusion is in contrast to prior work.

- However, we observe that BM25 could show a competitive ranking quality compared to TILDE and TILDEv2 which is in contrast to the findings about the relative performance of these three models on retrieval for short queries reported in prior work. This result raises the question about the use of contextualized term-based ranking models being beneficial in QBE setting. We follow-up on our findings by studying the score interpolation between the relevance score from TILDE (TILDEv2) and BM25.

不同 components 或方法性能可以累加

We see an accuracy increase of over 6 p.p. when fine-tuning the model and this is cumulative with RAG, which increases accuracy by 5 p.p. further. In one particular experiment, we also demonstrate that the fine-tuned model leverages information from across geographies to answer specific questions, increasing answer similarity from 47% to 72%.

跟前人不同

- StyleTTS 2 differs from its predecessor by modeling styles as a latent random variable through diffusion models to generate the most suitable style for the text without requiring reference speech, achieving efficient latent diffusion while benefiting from the diverse speech synthesis offered by diffusion models.
-

没有指出具体前人的不足

- However, the quest for robust and accessible human-level TTS synthesis remains an ongoing challenge because there is still room for improvement in terms of diverse and expressive speech [5, 6], robustness for out-of-distribution (OOD) texts [7], and the requirements of massive datasets for high-performing zero-shot TTS systems [8].

- However, our focus is more on how to better process and present data based on human preference, rather than merely retrieving it from databases. Additionally, while SQL is convenient, it can not directly satisfy common data analysis needs such as prediction and visualization.

absent of something 没人做过

- Absent of a clear benchmark for evaluating the performance of LLM routers, progress in this area has been hampered. To bridge this gap, we present RouterBench, a novel evaluation framework designed to systematically assess the efficacy of LLM routing systems, along with a comprehensive dataset comprising over 405k inference outcomes from representative LLMs to support the development of routing strategies. We further propose a theoretical framework for LLM routing, which provides a principled explanation for the observed performance differences between routing systems.
- Yet, the absence of a standardized benchmark for evaluating the performance of LLM routers hinders progress in this area. To bridge this gap, we present RouterBench, a novel evaluation framework designed to systematically assess the efficacy of LLM routing systems, along with a comprehensive dataset comprising over 405k inference outcomes from representative LLMs to support the development of routing strategies. We further propose a theoretical framework for LLM routing,

提出要探讨的问题

- We raise the question to what extent LLMs are capable of handling these applications off-the-shelf, i.e. without finetuning.

present 结论

- While there may be room for improvement through prompt engineering, our results aim to show the out-of-the-box LLM capabilities.

present algorithm

Algorithm 1: LLM k-shot predictions

Input: $\mathbf{x}$ – an instance from the training set
Input: $k(< M)$ – number of examples (max M)
Output: Δ_p – Softmax posteriors
begin
  $N_k(\mathbf{x}) \leftarrow \{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$
  `Instruction` ← "Predict the type of $\langle \mathbf{x} \rangle$ as one of $\{\langle C_0 \rangle, \ldots, \langle C_{p-1} \rangle\}$ given the following example".
  for $i \leftarrow 1$ *to* k **do**
    `Instruction.append`("Example: $\langle \mathbf{z}_i \rangle$ is a representative of class $\langle y(\mathbf{z}_i) \rangle$")
  $\Delta_p \leftarrow$ `LLM(Instruction)`
  return Δ_p

Algorithm 2: Optimal number of examples

Input: $\mathcal{T}$ – a training set of labelled instances
Output: $\mathcal{K} = \cup_{\mathbf{x} \in \mathcal{T}} k^*(\mathbf{x})$ – Number of examples yielding the most confident and correct predictions for each instance $\mathbf{x} \in \mathcal{T}$
begin
  for $\mathbf{x} \in \mathcal{T}$ **do**
    `max_confidence` $\leftarrow 0$; $k^* \leftarrow 1$
    for $j \leftarrow 0$ *to* M **do**
      $\Delta_p \leftarrow$ `LLM` k-`shot predictions`$(\mathbf{x}, j)$ `// Call Algorithm 1, i.e., try to predict with` j `examples`
      $\hat{y}(\mathbf{x}) \leftarrow \text{argmax}\Delta_p$ `// Get the predicted class`
      `confidence`$\leftarrow \Delta_{\hat{y}(\mathbf{x})} \mathbb{I}(\hat{y}(\mathbf{x}) = y(\mathbf{x}))$ `// Check if the predicted class is the correct one and record the prediction confidence`
      if `confidence > max_confidence` **then**
        `max_confidence` $\leftarrow$ `confidence // Keep track of the least uncertain correct prediction`
        $k^* \leftarrow j$
    $\mathcal{K} \leftarrow \mathcal{K} \cup k^*$
  return $\mathcal{K}$

can potentially be found at the very top ranks, as a result of which, a small number of examples should potentially work well. On the other hand, a low QPP estimate likely indicates that the very top ranked examples are not likely to be useful for downstream prediction, in which case it should be better to employ a large number of examples. This approach of selecting rank cutoffs (with an upper bound) as a function of the QPP scores has been applied to determine a variable depth of relevance assessments required for a robust retrieval evaluation [25].

by θ, via optimising:

$$\text{argmin}_\theta \sum_{\mathbf{x} \in \mathcal{T}, k^* \in \mathcal{K}} \mathcal{L}(\mathbf{x}^\mathrm{T}\theta, k^*), \tag{3}$$

小的优化，集成创新

- Retrieve Anything To Augment Large Language Models
  - Training such a unified model is non-trivial, as various retrieval tasks aim to capture distinct semantic relationships, often subject to mutual interference. To address this challenge, we systematically optimize our training methodology. This includes reward formulation based on

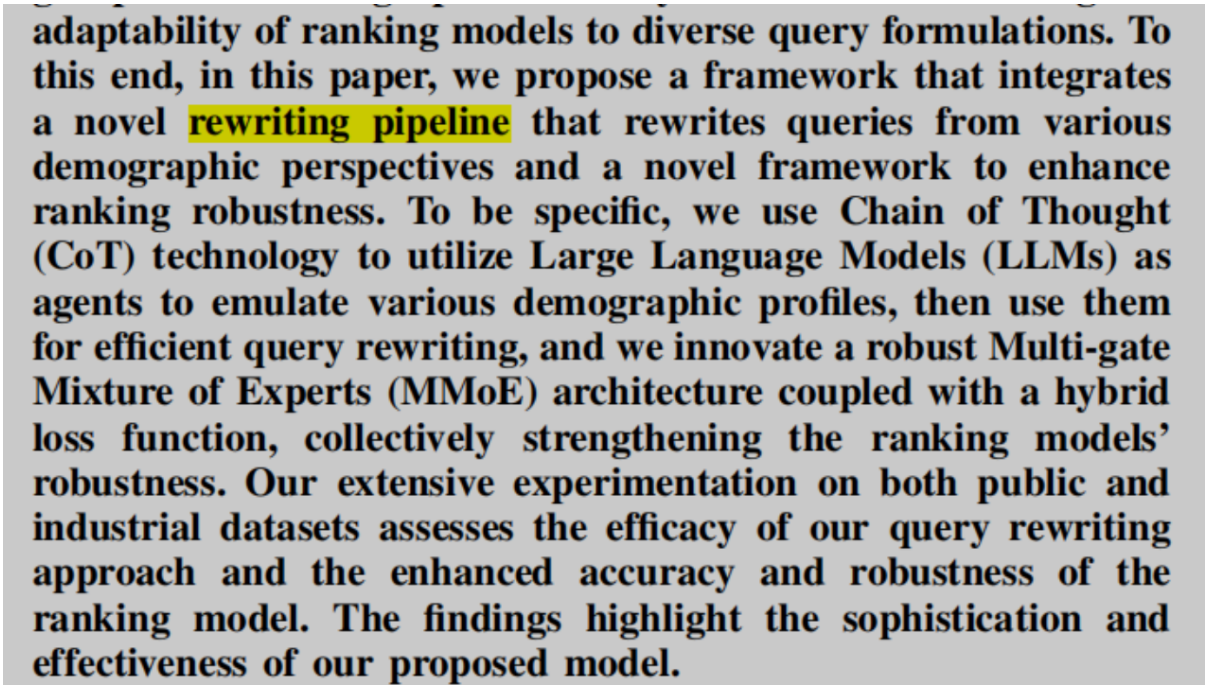
LLMs' feedback, the stabilization of knowledge distillation, multi-task fine-tuning with explicit instructions, and homogeneous in-batch negative sampling.

others

- The challenges faced by RAG systems, such as ensuring contextually appropriate and up-to-date data, are addressed by the dynamic nature of knowledge graphs.
-

Paper Structure

Abstract

- 1. 研究的问题是什么，及重要性。
- 2. 当前主流的方法（特别是你要比较的方法）是什么，存在那些问题。
- 3. 你是如何创造性的解决了这些问题的，具体怎么做的。
  - 
    adaptability of ranking models to diverse query formulations. To this end, in this paper, we propose a framework that integrates a novel rewriting pipeline that rewrites queries from various demographic perspectives and a novel framework to enhance ranking robustness. To be specific, we use Chain of Thought (CoT) technology to utilize Large Language Models (LLMs) as agents to emulate various demographic profiles, then use them for efficient query rewriting, and we innovate a robust Multi-gate Mixture of Experts (MMoE) architecture coupled with a hybrid loss function, collectively strengthening the ranking models' robustness. Our extensive experimentation on both public and industrial datasets assesses the efficacy of our query rewriting approach and the enhanced accuracy and robustness of the ranking model. The findings highlight the sophistication and effectiveness of our proposed model.
- 4. 实验结果是什么？结论是什么？有什么重要/有趣的发现。
  - 例如在 xx 最有代表性的数据集上，比 SOTA 在那些点上还好。
  - 个人偏向：结果呈现时直接给出具体数据。例如：We use a test set annotated by academic researchers in the fields of quantum physics and computer vision to evaluate our system's performance. The results show that DocReLM achieves a Top 10 accuracy of 44.12% in computer vision, compared to Google Scholar's 15.69%, and an increase to 36.21% in quantum physics, while that of Google Scholar is 12.96%.
  -

引出你做的事

- However, despite the success of foundation models in modalities such as natural language processing and computer vision, the development of foundation models for time series forecasting has lagged behind. We present Lag-Llama, a general-purpose foundation model for univariate probabilistic time series forecasting based on a decoder-only transformer architecture that uses lags as covariates.

Introduction

- 研究的问题是什么，为什么重要。
- 回顾最主要的对论文研究问题的主要研究工作、进展，
- 前任研究工作的主要不足在哪里，为什么到现在还没人解决难点在哪里？（如果没有难点，那你的研究工作也没意义？）|或者你发现了什么研究空白，确立研究机会。
    - 从普遍的 problem 到具体的 question
- 针对以上问题，你的研究思路是什么？你是如何创造性的解决这个问题的？
- 你的方法效果怎么样？实验结果是什么？
- 总结该文的主要贡献

如何说现有工作的缺点

- SPACE-3
- Despite the remarkable progress of previous PCMs on dialog understanding or dialog generation, there are still several technical challenges to constructing an effective and unified pre-trained conversation model. **First**, most current PCMs are tailored for one specific task like dialog understanding or dialog generation, as illustrated in Figure 1 (a)-(b). Limited exploration has been attempted to solve the three sub-tasks jointly in a unified framework. For example, although TOD-BERT significantly improves the performance of a wide range of dialog understanding tasks [99], it is difficult to apply TOD-BERT for dialog generation due to its bidirectional nature [92]. **Second**, a common idea of existing PCMs is to directly train existing PLMs on dialog corpora with vanilla language model objectives while neglecting the task-flow characteristics in TOD. For example, PPTOD [88] proposes to pre-train a dialog model by primitively amalgamating different dialog tasks with heterogeneous annotations into a text-to-text format like T5, which decreases the performance on individual TOD tasks. Different from open-domain chatting bots, TOD systems aim to help users accomplish certain tasks with a controllable step-by-step procedure, therefore it is essential to explicitly incorporate the task-flow into pre-training for fully exploiting task-oriented dialog features. **Third**, in previous PCMs, leveraging semantic structures and manual annotations (e.g., intents or slots) of dialogs to learn better pre-trained dialog representations still remains unexplored. Nevertheless, there have been many works demonstrating that labeled data can not only

samples

- Top-Down Partitioning for Efficient List-Wise Ranking [Abstract]. Argument 清晰具体
    - Large Language Models (LLMs) have significantly impacted many facets of natural language processing and information retrieval. Unlike previous encoder-based approaches, the enlarged context window of these generative models allows for ranking multiple documents at once, commonly called list-wise ranking. However, there are still limits to the number of documents that can be ranked in a single inference of the model, leading to the broad adoption of a sliding window approach to identify the k most relevant items in a ranked list. We argue that the sliding window approach is not well-suited for list-wise re-ranking because it (1) cannot be parallelized in its current form, (2) leads to redundant computational steps repeatedly re-scoring the best set of documents as it works its way up the initial ranking, and (3) prioritizes the lowest-ranked documents for scoring rather than the highest-ranked documents by taking a bottom-up approach. Motivated by these shortcomings and an initial study that shows list-wise rankers are biased towards relevant documents at the start of their context window, we propose a novel algorithm that partitions a ranking to depth k and processes documents top-down. Unlike sliding window approaches, our algorithm is inherently parallelizable due to the use of a pivot element, which can be compared to documents down to an arbitrary depth concurrently. In doing so, we reduce the number of expected inference calls by around 33% when ranking at depth 100 while matching the performance of prior approaches across multiple strong re-rankers.

Related Work

- 1. 深入地分类介绍了相关的工作：
    - 2. 介绍时，要说清楚跟当前论文的关系。
    - 3.

Method

- 引出部分（最好有个整体架构/流程图）： Figure 2 presents an overview of our approach for open-domain retrieval,

好的参考

- StyleTTS：
    - 先写一个 A. Proposed Framework 把问题和符号定义清楚。

2 Method

Our approach follows the *text-to-text* framework (Raffel et al., 2019). This means that all the tasks are framed as follows: the system gets a *text query* as input, and generates a *text output*. For example, in the case of question answering, the query corresponds to the question and the model needs to generate the answer. In the case of classification tasks, the query corresponds to the textual input, and the model generates the lexicalized class label, i.e. the word corresponding to the label. We give more examples of downstream tasks, from the KILT benchmark in Figure 2. As many natural language processing tasks require *knowledge*, our goal is to enhance standard text-to-text models with retrieval, which, as we hypothesise in the introduction, may be crucial to endow models with few-shot capabilities.

2.1 Architecture

Our model is based on two sub-models: the *retriever* and the *language model*. When performing a task, from question answering to generating Wikipedia articles, our model starts by retrieving the top-k relevant documents from a *large corpus* of text with the retriever. Then, these documents are fed to the language model, along with the query, which in turns generates the output. Both the retriever and the language model are based on pre-trained transformer networks, which we describe in more detail below.

Retriever. Our retriever module is based on the Contriever (Izacard et al., 2022), an information retrieval technique based on continuous dense embeddings. The Contriever uses a dual-encoder architecture, where the query and documents are embedded independently by a transformer encoder (Huang et al., 2013; Karpukhin et al., 2020). Average pooling is applied over the outputs of the last layer to obtain one vector representation per query or document. A similarity score between the query and each document is then obtained by computing the dot product between their corresponding embeddings. The Contriever model is pre-trained using the MoCo contrastive loss (He et al., 2020), and uses unsupervised data only. As shown in the following section, an advantage of dense retrievers is that both query and document encoders can be trained without document annotation, using standard techniques such as gradient descent and distillation.

- are llms all you need for TOD

3 Method

We introduce our method step-by-step. An overall description of the proposed pipeline is shown in Figure 2. The system consists of a pretrained LLM and an (optional) context store in a vector database. Three LLM calls are performed in each dialogue turn, with specific prompts (see Section 3.1). First, the LLM performs domain detection and state tracking (Section 3.2). The updated belief state informs a database query, whose results are used in the subsequent LLM-based response generation step (Section 3.3). In the few-shot setting, the context store is used to store a limited number of examples from the training set, which are retrieved based on similarity with the conversation context and included in LLM prompts (see Section 3.4).

- itransformer

Experiments

Compared Methods

- We evaluate Mimir against the following methods:

SMT+BM25: This method uses Statistical Machine Translation (SMT) to change queries from one language to English. It builds a translation table for each language pair using the GIZA++ toolkit, picking the top 10 translations for each query term. These terms are then combined into a new query using Galago's #combine function. Finally, BM25 is used to find documents based on these translated queries.

NMT+BM25: This approach improves on SMT by using Neural Machine Translation (NMT), which generally translates better. It translates queries into English with an NMT model, then uses BM25 to retrieve documents based on these translations.

Code-Switch: This technique uses data augmentation to better prepare for cross-lingual tasks. It applies a code-switching method to the queries in the MS MARCO dataset, then uses these queries to train the ColBERT retrieval model.

LaBSE: The *Retriever* part of MIMIR starts with LaBSE [9], a model known for its language understanding. We evaluate the *Retriever* using CLIR data without any prior training specifically for this task, known as a zero-shot setting.

Translate-Test: Similar to the NMT+BM25 method, this strategy involves translating the query into English using a Neural Machine Translation (NMT) model. After translation, the query is matched against English documents using a monolingual neural retrieval model like ColBERT, focusing solely on English-to-English comparisons.

BLOOMZ-7B1: The BLOOMZ model, introduced by Muennighoff et al. [36], is a versatile multitask model that's been fine-tuned on BLOOM, a base model known for its proficiency in 46 languages. For our experiments, we use the 7.1 billion parameter version of BLOOMZ, specifically after preparing it with data from the MS MARCO dataset.

GPT-3.5-TURBO: GPT-3.5-TURBO stands out as a leading large language model (LLM), benefiting from advanced tuning techniques,

Discussion

- 1. 描述实验结果
  - 提供必要的数据和数量，确保准确性；
  - 使用图标、表格或图像可视化工具来呈现
- 2. 解释结果
  - 说明你观察到的现象或趋势，并**解释**
  - 结果与最初假设是否有差异
  - 进行统计分析

Conclusion

- 一总结、一结果、一展望
- 总结做了什么工作（1-2 句）

- 主要结果是什么
- 展望工作的重要性，升华研究意义

paper style examples

good in general

- Query in Your Tongue: Reinforce Large Language Models with
    - good in general
- Optimizing Error-Bounded Lossy Compression for Scientific Data on GPUs

comparable studies

survey

- Dense Text Retrieval based on Pretrained Language Models: A Survey

evaluation only

- More Room for Language: Investigating the Effect of Retrieval on Language Models
    - examine how retrieval augmentation affects the behavior of the underlying language model.

no innovation but vertical application

- BMRETRIEVER: Tuning Large Language Models as Better Biomedical Text Retrievers

Tables

color and title

Model	WORLD KNOWLEDGE			SYNTACTIC KNOWLEDGE				LANGUAGE UNDERSTANDING		
	Concept Net (MRR ↑)	SQuAD (MRR ↑)	TREx (MRR ↑)	linear probing (LAS ↑)	attention probing (UUAS ↑)	BLiMP (Acc. ↑)	MSGS (LBS ↑)	LAMBADA (Acc. ↑)	GLUE (Avg. ↑)	SQuAD (F_1 ↑)
REFERENCE MODEL (110M)										
bert-base-cased	*26.0*	*34.0*	*62.0*	*82.0*	*45.1*	*85.6*	*-0.10*	*44.8*	*82.1*	*88.4*
BASE (98M)										
− retrieval	**20.3**	**32.1**	**53.6**	78.1	48.0	82.9	-0.47	**46.0**	**82.2**	**91.2**
+ retrieval (50% noise)	17.7	23.2	49.1	79.8	51.3	81.3	**-0.37**	43.2	82.0	90.7
+ retrieval (25% noise)	18.1	23.4	48.3	79.9	51.6	82.7	-0.38	40.6	81.9	90.2
+ retrieval (0% noise)	14.9	15.8	41.5	**80.2**	**51.8**	**83.2**	**-0.37**	37.5	81.2	89.7
SMALL (28M)										
− retrieval	**17.2**	**28.3**	**47.4**	71.2	49.7	78.6	-0.56	**35.1**	78.0	**88.6**
+ retrieval	11.8	15.3	36.3	**71.7**	**50.4**	**78.8**	**-0.53**	26.2	**78.4**	86.2
X-SMALL (9M)										
− retrieval	**9.9**	**14.7**	**39.2**	63.3	45.5	**73.4**	**-0.55**	**25.3**	75.2	**81.1**
+ retrieval	7.5	10.6	23.4	**63.6**	**49.2**	73.3	-0.57	19.3	**76.0**	78.7

Table 1: The overall evaluation scores for all sets of tasks, are divided into three categories. + denotes models pretrained with retrieval augmentation while − denotes standard models pretrained without retriever; note that the evaluation is done without any retrieval mechanism for all models (see Section 4). We divide the models into three subsets based on their size and also give the reference scores of the official `bert-base-cased` model evaluated with our pipeline. We highlight the best results for each model size in **boldface** and measure the average score across 5 runs, when applicable. The red color indicates worse results than the no-retrieval baseline and vice-versa for the blue color.

-
- from: More Room for Language: Investigating the Effect of Retrieval on Language Models

tools

polish

- https://www.citexs.com/Editing