LLMAEL: Large Language Models are Good Context Augmenters for Entity Linking

Anonymous ACL submission

Abstract

Entity Linking (EL) models are well-trained at mapping mentions to their corresponding entities according to a given context. However, EL models struggle to disambiguate long-tail entities due to their limited training data. Meanwhile, large language models (LLMs) are more robust at interpreting uncommon mentions. Yet, due to a lack of specialized training, LLMs suffer at generating correct entity IDs. Furthermore, training an LLM to perform EL is cost-intensive. Building upon these insights, we introduce LLM-Augmented Entity Linking (LLMAEL), a plug-and-play approach to enhance entity linking through LLM data augmentation. We leverage LLMs as knowledgeable context augmenters, generating mention-centered descriptions as additional input, while preserving traditional EL models for task specific processing. Experiments on 6 standard datasets show that the vanilla LLMAEL outperforms baseline EL models in most cases, while the fine-tuned LLMAEL set the new state-of-the-art results across all 6 benchmarks.

1 Introduction

Entity linking (EL) establishes connections between mentions in textual contexts and entities in a target knowledge base (KB). It plays an important role in many applications requiring semantic understanding, such as question answering (Yao et al., 2023; Perez-Beltrachini et al., 2023; Xu et al., 2023), dialogue generation (Cui et al., 2022; Rückert et al., 2022), and making recommendations (Wang et al., 2022; Balloccu et al., 2022).

However, EL is still a challenging task as it requires two distinct capabilities: **(a) Task Specification**, which encompasses a thorough understanding of the entity linking task and the precise requirement for its output format, and **(b) Entity Knowledge**, which involves the possession of substantial knowledge about the target entity. Trained specifically for EL, traditional EL models (Wu et al., 2020; Cao et al., 2021; Ayoola et al., 2022) excel in task specification, capable of producing results that exactly satisfy the format requirement of the EL task. Meanwhile, extensively pre-trained large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023) are natural repositories of expansive world knowledge, possessing a vast reservoir of information pertinent to any given entity.

However, these two streams of models each present their own limitations for EL. Compared to knowledgeable LLMs, traditional EL models are constrained by their limited knowledge accumulated during model training, resulting in a narrower scope of knowledge about entities. While the training data for EL models equips them to master the specification of the entity linking task, it falls short of providing them with comprehensive knowledge of all entities, especially unpopular entities that scarcely or never appear in the training data.

Similarly, relying exclusively on LLMs for entity linking comes with its own set of drawbacks. With a primary design for language modeling, LLMs struggle to perform tasks that demand precise specifications (Peng et al., 2023). More specifically, entity linking requires the production of exact entity IDs within a knowledge base. The correct generation of these IDs, which differ fundamentally from natural language, poses a significant challenge to LLMs. Although LLMs can partially learn the task specification for generating entity IDs via in-context learning (Brown et al., 2020, ICL), we observe that LLMs tend to produce fictional entity IDs, which is recognized as hallucination (Rawte et al., 2023). This leads to erroneous linkage of mentions to non-existent KB entities.

To address the limitations inherent in traditional EL models and modern LLMs respectively, we design a novel pipeline method that capitalizes on the strengths of both approaches. We present LLM-Augmented Entity Linking (LLMAEL), a plug-and-play method to bolster entity linking through

LLM data augmentation. Instead of demanding LLMs to perform EL directly, we leverage LLMs as context enhancers, supplementing EL models with additional context regarding a specific mention. Our method consists of three primary stages: (1) context augmentation, where LLMs are prompted to augment the original mention-context pair by generating supplementary mention descriptions, (2) data fusion, where the LLM-augmented context is integrated into a selected EL model, and (3) EL execution, where the EL model is employed to retrieve the target entity.

LLMAEL enhances entity linking by integrating the broad world knowledge and text generation abilities of LLMs with the specialized KB interaction skills of EL models. First, we leverage LLMs for context augmentation, enriching mentions with LLMs' world knowledge while also condensing mention-related information from the provided context. Then, we employ an EL model to execute entity linking, thus minimizing the risk of obtaining invalid entity IDs due to LLM hallucination. Our method effectively combines the strengths of EL models and LLMs while addressing their respective shortcomings, leading to a more accurate and reliable EL solution.

For all 3 selected EL models, our vanilla LLMAEL achieves at least SOTA performance on more than 4 datasets. Our fine-tuned LLMAEL yields new SOTA results across all 6 datasets, achieving an average 1.24% accuracy gain. Furthermore, employing optional techniques such as context-joining and ensemble further boosts performance.

2 Preliminaries and Related Work

We give formal definition and notations for entity linking. We also introduce related work targeting entity linking, and methods using LLMs directly or as context augmenters for various tasks.

2.1 Task Definition

Entity Linking (EL) is the task of mapping mentions from a given context to KB entities. Formally, knowledge base G consists of the set of entities that are unique objects in the real world $\{e\}$. The input of entity linking is a textual context c, embedded with multiple entity mentions, denoted as $c = \ldots t_1||m_1||t_2||m_2||t_3 \ldots$, where t_i are textual spans and m_i are entity mentions. The goal of entity linking is to obtain a correct list of mention-entity pairs $\{(m_i, e_i)\}_{i\in[1,k]}$.

2.2 Related Work

Entity Linking. It has been a long-standing goal to develop reliable entity linking solutions. The most widely adopted procedure to tackle EL is a two-stage architecture (Sevgili et al., 2022), which divides EL into two sequential phases: candidate generation and entity re-ranking. Most models approach the candidate generation phase as a retrieval problem, aligning mentions to entities according to a specific metric (Wu et al., 2020; Logeswaran et al., 2019; Le and Titov, 2018). With the development of generative language models, it becomes possible to treat candidate generation as a text generation task (Cao et al., 2021), training the model to generate unique entity names in the knowledge base directly based on the contextual information.

Most recent works prove that concept information about mentions is useful for EL, thus fine-grained entity typing is also integrated as part of the pipeline and has been applied to numerous EL models (Ayoola et al., 2022; Onoe and Durrett, 2020; Raiman and Raiman, 2018). This suggests that augmenting mentions with additional information about the entities potentially facilitates the entity linking process.

LLMs as Executor for Downstream Tasks. In-context learning, or few-shot prompting, is a prevalent strategy that directs LLMs to perform specific tasks by providing them in-prompt demonstrations. With the outstanding accomplishments of LLMs like GPT-3 (Brown et al., 2020) and LLaMA2 (Touvron et al., 2023), LLMs have achieved impressive results in downstream tasks through in-context learning, including question answering, summarization, and machine translation, *etc*. However, LLMs still struggle when executing specification-heavy tasks (Peng et al., 2023), yielding results that are far from state-of-the-art. Hence, employing LLMs through in-context learning may not always be the best solution for any given task (*e.g.,* EL).

LLMs as Context Augmenters for Downstream Tasks. LLMs are primarily designed for context generation, which is their strongest advantage. Multiple studies have demonstrated that LLM-generated contexts present outstanding qualities, outperforming contexts obtained from information retrieval methods (Yu et al., 2022; Chen et al., 2023). Furthermore, compared to retrieved contexts, LLM-generated contexts contribute to

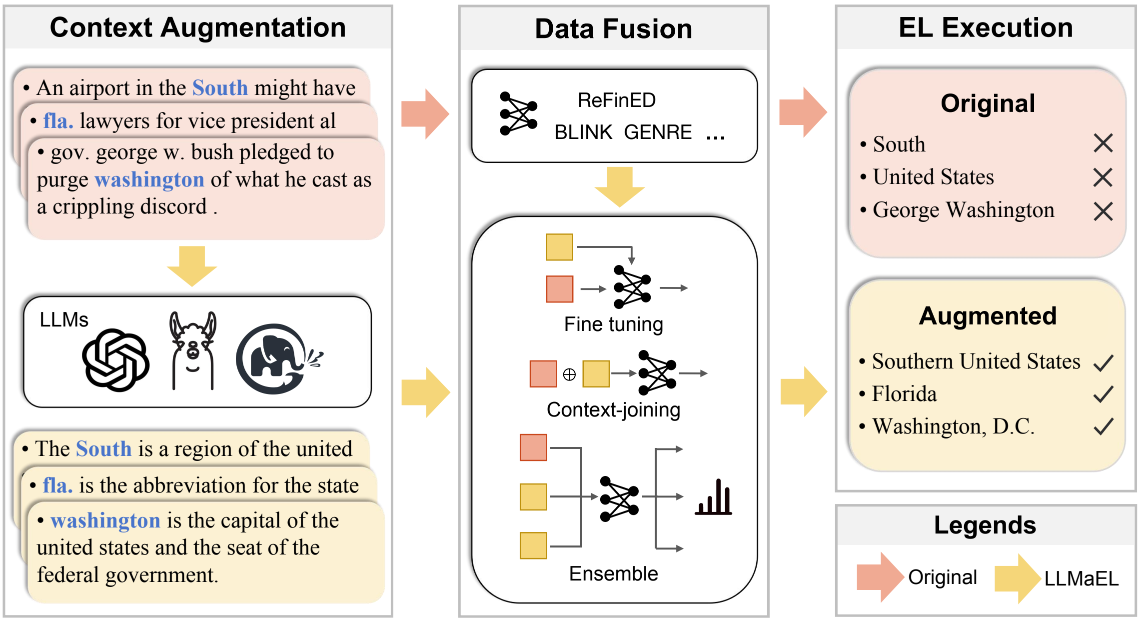

Figure 1: The overall architecture of our approach. We mark the traditional entity linking process in pink and our method in yellow. Mentions that need to be executed by entity linking are marked with blue.

better downstream task performance (Chen et al., 2023). With such insights, a bright solution is to leverage LLMs as context enhancers, generating contexts for downstream tasks as additional input. Liu et al. (2022) propose using LLM context augmentation for commonsense reasoning, achieving state-of-the-art results on multiple reasoning tasks. Similarly, Balkus and Yan (2022) improve text classification with GPT-3 augmented data, yielding higher consistent accuracy on unseen examples.

3 Methodology

LLMAEL is a plug-and-play enhancement method for entity linking using LLMs as context augmenters. It mainly includes three building blocks. (1) **Context augmentation** is the most basic element for LLMAEL, which effectively elicits LLMs to generate enriched context with more information for entity linking. (2) **Data fusion** designs multiple strategies to integrate the LLM-generated content with the original context, aiming to improve diverse off-the-shelf EL models. (3) **EL execution** finally conducts the entity linking task. Figure 1 illustrates the overall workflow of LLMAEL.

3.1 Context Augmentation

In a nutshell, we rely on decoder-only LLMs, which are used to map the original context along with an information-expansion instruction prompt into enhanced context. Formally, we denote LLM context generation as a function: $c' = \mathrm{LLM}(p, c, m_i)$, where p is a specially designed prompt to instruct LLMs to perform context augmentation. c' is the supplemented textual information for the ith mention m_i.

Prompt Design. The main strategy to control LLMs to augment context as expected is in-context learning (Brown et al., 2020), which effectively constraints the output format of LLMs. Thus, our prompt includes two parts: (1) task specifications for expanding information, and (2) exemplars of paired original contexts and LLM-generated descriptions.

For task specification, we use the following template to ask LLMs to complete the sentence:

```
Consider the following text.
Text: [CONTEXT]
Please provide me more descriptive
  information about [MENTION] from
  the text above.
Make sure to include [MENTION] in
  your description.
```

where [CONTEXT] and [MENTION] are placeholders to be filled before feeding into LLMs. It is worth noting that our instruction requires LLMs to mention the entity again in the augmented context, which provides flexibility for data fusion.

ID	Context Order	Mention Offset
0	LLM-only	LLM
1	LLM + Original	LLM
2	LLM + Original	Original
3	Original + LLM	LLM
4	Original + LLM	Original

Table 1: Five context-joining strategies of LLMAEL, enumerating the arrangements of the original and LLM-generated contexts in the final augmented context, which is to be inputted to the EL model. *Context Order* denotes the sequential order of the original and LLM-generated contexts. Since the same mention appears at least twice in the augmented context (one in the original context and one in the LLM-generated context), *Mention Offset* specifies the final mention-span to be provided to the EL model.

For in-prompt demonstrations, we bootstrap examples via zero-shot prompting. To ensure the quality of these demonstrations, we first generate a sufficiently large amount of contexts via zero-shot prompting, and then manually filter out high-quality completions. The final prompt's exemplars are then selected from this high-quality sample.

Due to the limited input size of LLMs, the final prompt includes three distinct examples. We show details of our prompt in Tables 6 and 7 in the appendix.

3.2 Data Fusion

Data fusion designs strategies from multiple perspectives to incorporate LLM-augmented context c'. In particular, **context-joining strategies** fuse c' within the original context c; **EL model fine-tuning** fuses the knowledge in c' into the EL model; and **ensemble** fuses multiple LLM-augmented contexts.

Context-Joining Strategies. The most direct way to integrate LLM-augmented context c' with original context c is to concatenate them and feed them into off-the-shelf EL models, which we denote as our *vanilla* implementation strategy for LLMAEL. To this end, we consider the following two design problems: (1) **Context order.** What sequence order should be adopted for concatenating the contexts c and c'? (2) **Mention offset.** Given that the mention occurs in both c and c', which context's spans should be utilized to refer to the mention when invoking EL models?

For LLMAEL, we design 5 potential context-joining strategies, as shown in Table 1. Joining strategy 0 uses LLM-generated c' as a direct substitute for c, while joining strategies 1 to 4 present all 4 possible combinations over the distinct orders of the two contexts and the choice of the mention offset. We empirically find that different EL models perform best under different joining strategies, so we maintain the choice of context-joining strategy as a hyper-parameter, providing space for user adjustment across different settings.

EL Model Fine-tuning. While the vanilla LLMAEL already demonstrates performance enhancements, the style and distribution of the augmented context are unfamiliar to EL models. To mitigate the gap between the data distribution that EL models are familiar with and the LLM-augmented contexts, we further fine-tune existing EL models. Specifically, we utilize existing EL training datasets and run the context augmentation step in Section 3.1, generating mention-centered descriptions for the entire dataset using an LLM. Then, we augment the training set with the generated descriptions using the optimal context-joining strategy for the selected EL model. Finally, we apply this augmented training set to fine-tune the EL model.

Ensemble. Inaccuracies in LLMAEL's performance may occur when the LLM generates wrong mention descriptions, misguiding the EL model to select a wrong entity. Hence, we also expand LLMAEL with ensemble techniques, attempting to improve our method's robustness through diversified sampling. We sample mention descriptions across multiple LLMs and evaluate the diversified samples through both hard-voting and soft-voting classifier methods.

3.3 EL Execution

In the final phase of EL execution, the EL model is employed to output the entity ID in the target knowledge base. Compared to directly tasking LLMs to perform entity linking, LLMAEL injects task specification knowledge using EL models. Additionally, it augments EL models with sufficient entity knowledge from the infused LLM data.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate LLMAEL on 6 standard EL datasets AIDA-YAGO2 (Hoffart et al., 2011), MSNBC (Cucerzan, 2007), AQUAINT (Milne and Witten, 2008), ACE04 (Ratinov et al., 2011), WNED-CWEB (Gabrilovich et al., 2013), and WNED-WIKI (Guo and Barbosa, 2018). The datasets are first augmented using our chosen LLM, then evaluated on our selected EL models.

Backbone Models for LLMAEL. For our main experiments, we use gpt-3.5-turbo-instruct (OpenAI, 2023) as our LLM, considering its good performance on text completion. As LLMAEL is a plug-and-play framework for any EL models, to implement LLMAEL, we select three most widely adopted EL models as our backbone: BLINK (Wu et al., 2020), a classical bi-encoder cross-encoder EL model; GENRE (Cao et al., 2021), an autoregressive generative EL solution; and ReFinED (Ayoola et al., 2022), an enhanced EL method using entity types and descriptions. To implement these models, we utilize their original implementations. That is, the full BLINK model[1], the fairseq-AIDA GENRE model[2], and the AIDA ReFinED model[3]. For unified implementation, we follow BLINK and ReFinED to execute GENRE without the candidate set.

Baselines. We compare LLMAEL with two categories of baselines: (1) **LLMs for EL.** We leverage gpt-3.5-turbo-instruct to execute the EL task directly. We provide the LLM with a few shot prompt that includes paired examples of mention contexts and gold entities. The concrete prompt is included in the Table 8 in the appendix. (2) **Traditional EL models.** We compare with each of our three backbone EL models to conduct EL on the original datasets, without any LLM data augmentation.

Evaluation Metrics. We use disambiguation accuracy as our evaluation metric. The unweighted macro average over all test sets is also reported.

The Unified Context-Joining Strategy. For all our implement LLMAEL variants, we apply the development (dev) subset of AIDA-YAGO2 to select the optimal context-joining strategy. In the main experiments detailed in Section 4.2.1, we adopt a unified strategy—strategy 4—that yields the highest average accuracy across all EL models. We hypothesize that this strategy outperforms others because most EL models are more adept at processing original contexts, thus performing better when LLM-generated contexts are placed towards the end. Interestingly, each EL model shows optimal performance with a different joining strategy. For EL models whose optimal test-time strategy diverges from the unified strategy, we present the outcomes achieved using their respective optimal strategies in Section 4.2.2.

Fine-tuning. We select our best-performing EL model ReFinED for model fine-tuning. We use the train and dev splits from the AIDA-YAGO2 dataset as our training and evaluation data. To avoid model over-fitting, we leverage ReFinED's wikipedia model[4] for fine-tuning. Specifically, we first employ gpt-3.5-turbo-instruct to augment the datasets under the model's optimal context-joining strategy, then apply the augmented datasets to the fine-tuning process.

4.2 Experimental Results

4.2.1 Main Results

We compare LLMAEL with baselines and report the results in Table 2. For LLMAEL, we evaluate the vanilla implementation of LLMAEL with optimal context-joining strategy searched from the dev set of AIDA-YAGO2. We further fine-tune LLMAEL with the best performing backbone EL model (ReFinED).

We find that even the vanilla implementation of LLMAEL uniformly brings performance gain, comparing to the average accuracy of their original backbone. For each datasets, all the implementations at least improve performance on more than 4 datasets, with LLMAEL $\times$ GENRE outperforming 5 out of 6 datasets with an average enhancement of 0.66%.

The fine-tuned LLMAEL surpasses all six benchmarks with an average 1.24% accuracy gain, yielding new state-of-the-art results over all six datasets. This supports our hypothesis that fine-tuning further amplifies our method's performance, as it better aligns EL models with the distribution characteristics of LLM-augmented contexts.

4.2.2 Ablations

Model-Specific Context-Joining Strategies. In Table 2, we employ the unified context-joining

[1]BLINK's full cross-encoder model

[2]The GENRE model developed using the fairseq toolkit and officially fine-tuned on AIDA-YAGO2

[3]The ReFinED model officially fine-tuned on AIDA-YAGO2

[4]The ReFinED model that is not officially fine-tuned on AIDA-YAGO2

Method	AIDA	MSNBC	AQUA	ACE04	CWEB	WIKI	AVG
LLM only	79.72	82.13	75.24	85.99	65.18	69.74	76.33
BLINK only	82.01	86.23	85.16	86.01	69.11	81.11	81.61
GENRE only	87.92	83.54	84.32	84.82	68.75	83.02	82.06
ReFinED only	92.25	87.10	87.53	87.75	72.96	85.18	85.46
LLMAEL × BLINK	81.60	86.56	85.16	86.01	69.30	81.06	81.62
LLMAEL × GENRE	87.83	85.37	85.14	84.82	70.63	83.10	82.82
LLMAEL × ReFinED	92.09	86.94	88.23	**88.14**	73.32	85.60	85.72
LLMAEL × $\text{ReFinED}_{\text{FT}}$	**92.38**	**88.63**	**89.47**	**88.14**	**75.09**	**86.48**	**86.70**

Table 2: Disambiguation accuracy scores across six test sets. The best value is in **bold** and second best is underlined. All models that involves BLINK, GENRE, or ReFinED are tested with scripts provided by each model's respective authors. AIDA refers to the test split of the AIDA-YAGO2 dataset. $\text{ReFinED}_{\text{FT}}$ corresponds to our customly fine-tuned version of ReFinED. The GENRE model is used without candidate sets.

Method	ID	AVG acc.
BLINK only	-	81.61
LLMAEL × BLINK	4	81.62
LLMAEL × BLINK	1*	**83.53**
GENRE only	-	82.06
LLMAEL × GENRE	4	82.82
LLMAEL × GENRE	2*	**83.03**

Table 3: Performance of the vanilla LLMAEL combined with GENRE and BLINK under the unified strategy 4 and each model's own optimal test-time strategy. The *ID* column refers to the selected context-joining strategy ID, while the *AVG acc.* column presents the unweighted macro average disambiguation accuracy score over all 6 test sets. The best average accuracy score for each model is in **bold**. Refer to Table 1 for detailed descriptions of all 5 joining strategies.

strategy, specifically strategy 4, chosen for its highest average accuracy score on the AIDA-dev dataset across all 3 EL models. Although strategy 4 proves to be the most effective for ReFinED during testing, it does not align as the optimal joining strategy for GENRE and BLINK.

Table 2 presents the optimal joining strategies for GENRE and BLINK at test time, alongside their respective average accuracy scores across all six datasets. Adopting these model-specific optimal strategies leads to a significant performance enhancement, with BLINK experiencing a substantial increase of 1.92% in accuracy.

Intriguingly, BLINK's optimal test-time strategy (strategy 1) has complete opposite parameters as unified strategy 4. We hypothesize that the reliance on AIDA-dev for selecting the optimal joining strategy might be a contributing factor. Given that BLINK, unlike the other two models, is not fine-tuned on the AIDA dataset, it may not resonate well with the textual distributions of AIDA datasets. Consequently, BLINK's performance on the AIDA-dev dataset does not accurately reflect its true preferences and capabilities.

Choosing Among LLMs. LLMAEL is adaptable to any LLM. In this section, we implement LLMAEL using two other widely recognized LLMs, namely Llama2-13b-chat (Touvron et al., 2023) and GLM-4 (Du et al., 2022). Table 4 presents our results in the upper half labeled *Single*.

For the vanilla LLMAEL (*ie.,* ReFinED), all three LLMs demonstrate an average performance enhancement. Among them, Llama2-13b-chat yields the most significant overall improvement, achieving an average accuracy of 87.28%. GLM-4 and gpt-3.5-turbo-instruct exhibit comparable performance, with a minimal 0.03% difference in their average accuracies.

For the fine-tuned LLMAEL (i.e., $\text{ReFinED}_{\text{FT}}$), gpt-3.5-turbo-instruct and GLM-4 demonstrate comparable levels of enhancement, while Llama2-13b-chat shows slightly weaker performance. This variation is caused by the differences in data formats produced by each LLM. Notably, when given identical instructions, only Llama2-13b-chat tends to start with a colloquial response to the instruction, generating phrases like *"Of course! Here are more details about..."* This characteristic negatively impacts the performance of EL models fine-tuned on data generated by gpt-3.5-turbo-instruct, leading to a degradation in performance.

It is also noteworthy that the performance of $\text{ReFinED}_{\text{FT}}$ also shows a significant enhancement

	EL Model	LLM(s)	MSNBC	AQUA	ACE04	WIKI	AVG
Single	ReFinED	-	87.10	87.53	87.75	85.18	86.89
		gpt-3.5-turbo-instruct	86.94	88.23	**88.14**	85.60	87.23
		Llama2-13b-chat	87.25	87.95	**88.14**	85.77	87.28
		GLM-4	86.94	87.95	**88.14**	85.75	87.20
	ReFinED$_{FT}$	-	**88.94**	89.34	**88.14**	85.99	88.10
		gpt-3.5-turbo-instruct	88.63	**89.47**	**88.14**	86.48	88.18
		Llama2-13b-chat	88.63	89.20	**88.14**	86.37	88.09
		GLM-4	88.63	89.34	**88.14**	86.51	88.16
Multi	ReFinED	Hard-voting ensemble	87.10	87.95	**88.14**	85.87	87.73
		Soft-voting ensemble	86.94	87.95	**88.14**	85.75	87.20
	ReFinED$_{FT}$	Hard-voting ensemble	88.79	89.20	**88.14**	**86.67**	**88.20**
		Soft-voting ensemble	88.63	**89.47**	**88.14**	86.56	**88.20**

Table 4: Disambiguation accuracy scores across four selected datasets, where LLMAEL applies different LLMs and ensemble techniques. The *Single* portion presents the EL model's performance on its own and after its integration with three individual LLMs. The *Multi* portion presents results obtained by the ensemble of all four outputs. The best value for each dataset is in **bold**.

when applied to the original contexts, registering an average performance improvement of 1.21%.

Ensemble. We use both hard-voting and soft-voting classifiers to perform ensemble. The hard-voting classifier is executed by selecting the most frequent outcome among multiple independently-generated results. In instances where multiple results share an equivalent frequency, the result with the highest probability level is selected. Conversely, the soft-voting classifier selects the final answer by aggregating the probabilities of all outcomes.

The *Multi* half of Table 4 illustrates our ensemble results. Both ReFinED and ReFinED$_{FT}$ are improved by the implementation of ensemble techniques. For ReFinED, ensemble using the hard-voting classifier achieves a highest average accuracy of 87.73%. This accuracy score is higher than the score obtained by the soft-voting classifier, and also higher than any one of ReFinED's 4 individual scores. This is because the hard-voting classifier is particularly effective when the performance of individual models are diverse. For most datasets, the original ReFinED model yields results that are apparently different to the other 3 LLM-enhanced models, contributing to the diversity of model performance. Meanwhile, when the performance of single models is relatively uniform, both ensemble methods—hard and soft-voting classifiers—exhibit comparable effectiveness. This phenomenon is evident in the performance outcomes of the ReFinED$_{FT}$, where the hard and soft-voting classifiers present equal average accuracies.

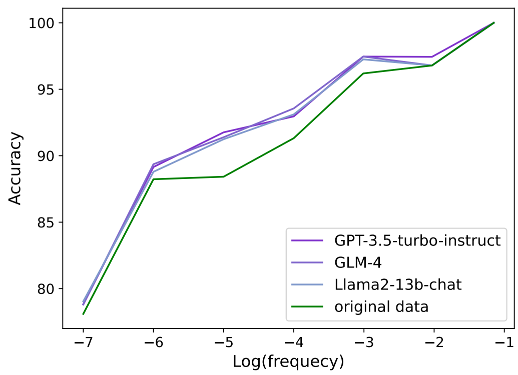

Figure 2: EL performance across entities of different frequencies. The green line illustrates the performance of the original ReFinED model applied to the original datasets. The purple lines illustrate the performance of our customly fine-tuned ReFinED$_{FT}$ model using LLM-augmented datasets.

4.3 Discussions

We delve deeper into LLMAEL by examining the following two discussion questions.

4.3.1 Does LLM-Augmented Data Improve EL Performance Over Long-Tail Entities?

Entities vary in frequency, depending on how often they are referenced in the real world. EL models tend to perform better on high-frequency entities and poorly on low-frequency entities due to their limited training data.

LLMs possess more entity knowledge compared

Method	AIDA	MSNBC	AQUA	ACE04
BLINK only	82.01	86.23	85.16	86.01
LLM only	79.72	82.13	75.24	85.99
Re-rank-100	55.88	64.69	46.99	61.11
Re-rank-10	70.35	78.04	61.09	74.29

Table 5: Performance of BLINK and three configurations of LLM executing EL tasks, where the LLM employ the gpt-3.5-turbo-instruct model.

to EL models, which can be transferred to EL models through LLM data augmentation. Hence, we hypothesize that a core contributor to LLMAEL's effectiveness is its ability to enhance EL models over long-tail entities—entities that possess low frequencies.

To investigate this, we evaluate LLMAEL across entities of different frequencies. We select four datasets—MSNBC, AQUAINT, ACE04, and WNED-WIKI—to calculate the cross-dataset accuracy of each contained gold entity. For each entity, we assign its corresponding PageRank value from Wikidata5M (Wang et al., 2021) as its frequency. To simplify visualization, we normalize the frequencies using a base-10 logarithmic scale. Finally, we categorize all entities into seven buckets, each bucket comprising the entities that share the same integer part in their normalized frequencies.

Our findings are illustrated in Figure 2, where the horizontal axis presents normalized entity frequencies and the vertical axis presents the average accuracies of each entity bucket. LLMAEL improves the accuracy of entities with mid-to-low frequencies within the range of 10^{-6} to 10^{-2}, and refines the accuracy of entities with extremely low frequencies in the range of 10^{-7} to 10^{-6}.

Such results align with our hypothesis that LLM data augmentation enhances EL performance for long-tail entities. Furthermore, the results indicate that LLMs also improve performance for mid-tail entities. This improvement is likely because the LLM-generated data offers condensed mention information, thereby reducing the noise present in the original contexts.

4.3.2 Is There a Better Way to Leverage LLMs for EL?

Considering that many EL models, such as BLINK, operate by first retrieving candidate entities and then re-ranking them, a practical approach is to use EL models for candidate retrieval and LLMs for re-ranking. In this section, we explore whether this is a more effective way to leverage LLMs for EL.

We use BLINK's bi-encoder for candidate retrieval and gpt-3.5-turbo-instruct for re-ranking. We establish two re-ranking settings:

- **Re-rank-100:** Extract the top 100 candidate entities of BLINK's bi-encoder and task the LLM to select the final entity.

- **Re-rank-10:** Extract the top 10 candidate entities of BLINK's bi-encoder and augment each candidate with its Wikipedia abstract. The LLM selects the final entity using the abstracts as supplementary information.

As shown in Table 5, applying LLMs for entity re-ranking does not enhance EL performance; in fact, its efficacy is even lower than directly leveraging LLMs for EL. We observe two primary reasons for this inefficacy. First, the presence of similar candidate names confuses the LLM. Unlike demanding LLMs to directly generate entity names for mentions, asking LLMs to perform re-ranking requires them to discern the subtle distinctions among candidates. As highlighted by Peng et al. (2023), LLMs struggle to understand and distinguish complex contexts, leading to diminished performance. Secondly, presenting the LLM with multiple candidates often causes it to spread its focus across the entire context rather than concentrating on the specific mention. This results in the LLM to prioritize information that is distant and unrelated to the mention.

The relatively improved performance of the second re-ranking setting can be attributed to the reduced number of candidates and the inclusion of abstracts. These modifications make it easier for the LLM to understand and differentiate the candidate entities.

5 Conclusion

This paper presents LLMAEL, a lightweight and flexible pipeline approach to enhance entity linking through LLM data augmentation. It leverages the strengths of both EL models and LLMs with minimal costs, yielding promising results without the need for further training. Furthermore, it offers advanced data fusion options. For future work, we are interested in exploring more effective ways to integrate LLMs into entity linking, aiming to infuse fresh momentum to the field in the LLM era.

Ethical Considerations

Entity linking is a well established task, aiming to bridge textual data and structural data (*e.g.,* knowledge base). This work follows this setting, aiming to provide a better EL method with higher accuracy. As the proposed methodology LLMAEL is our main contribution, we are hereby to discuss potential misuse of LLMAEL.

Potential Misuse. The risk to misuse LLMAEL is the same as all other EL models, such as using entity linking models to decorate generated fake contents with apparently right but actually wrong reference. Moreover, we would like to point out that, as LLMAEL allow for utilizing a third party LLM to augment EL data. If the used LLM is jailbreaked or hacked to produce misinformation, it would result in cascading failure of LLMAEL.

Possible Biases. LLMs carry potential risks of generating biased or harmful content. Since our approach relies on LLMs for context generation, our pipeline method and fine-tuned model could inherit existing biases present in the LLMs' model weights.

Environmental Impact. The model inference and EL model fine-tuning phases of LLMAEL lead to energy and carbon costs. However, compared to methods that leverage LLMs through LLM fine-tuning, our method requires less energy expenses.

Limitations

As a pipeline method, LLMAEL relies heavily on the abilities of its selected EL models and LLMs. Yet, the most advanced LLMs currently available are commercial products, which incur costs for each API call. Furthermore, some LLMs show accessibility constraints. For instance, GPT-4 is not included in our experiments due to our limited access to the model.

References

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Salvador V Balkus and Donghui Yan. 2022. Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, pages 1–30.

Giacomo Balloccu, Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2022. Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. *ArXiv preprint*, abs/2310.07289.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Wen Cui, Leanne Rolston, Marilyn Walker, and Beth Ann Hockey. 2022. OpenEL: An annotated corpus for entity linking and discourse in open domain dialogue. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2245–2256, Marseille, France. European Language Resources Association.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Evgeniy Gabrilovich, Michael Ringgaard, Amarnag Subramanya, et al. 2013. Facc1: Freebase annotation of clueweb corpora.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8576–8583. AAAI Press.

OpenAI. 2023. Gpt-3.5 turbo. `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`. Accessed: 2024-02-09.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *ArXiv preprint*, abs/2311.08993.

Laura Perez-Beltrachini, Parag Jain, Emilio Monti, and Mirella Lapata. 2023. Semantic parsing for conversational question answering over knowledge graphs. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2507–2522, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: Multilingual entity linking by neural type system evolution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5406–5413. AAAI Press.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *ArXiv preprint*, abs/2309.05922.

Ulrich Rückert, Srinivas Sunkara, Abhinav Rastogi, Sushant Prakash, and Pranav Khaitan. 2022. A unified approach to entity-centric context tracking in social conversations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1275–1285, Marseille, France. European Language Resources Association.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Silei Xu, Shicheng Liu, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina Semnani, and Monica Lam. 2023. Fine-tuned LLMs know more, hallucinate less with few-shot sequence-to-sequence semantic parsing over Wikidata. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li, and Lei Hou. 2023. KoRC: Knowledge oriented reading comprehension benchmark for deep text understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *ArXiv preprint*, abs/2209.10063.

A Reproducibility Details

Datasets. For each EL model, we download their official training and testing datasets from their respective github repositories[5][6][7]. Each model provides an official version of our 8 selected datasets[8]. Since ReFinED is the only model that supports NIL entities, ReFinED's official 8 datasets contain the largest number of entries. Hence, we obtain LLM-generated context for each of ReFinED's 8 datasets. Then, we map these generated contexts to the official datasets of the other two EL models, ensuring correct alignment for each data entry.

Context-Joining. For joining strategies that incorporate context combination, contexts are merged using a newline symbol "\n". In the case of BLINK and ReFinED, contexts are fully combined in the specified order, without any truncation applied. In the case of GENRE, the contexts are first fully combined, then trimmed to the model's maximum input sequence length.

Testing Scripts. For all experiments that incorporate an EL model, we augment the model's official datasets with LLM-generated context, then use the model's official testing script to produce our results.

LLM Hyperparameters. For all LLM-generated responses, we set the number of maximum tokens to 150. We set the temperature of gpt-3.5-turbo-instruct to 0 and the temperature of Llama2-13b-chat and GLM-4 to 0.01. We adapt default values for all other hyperparameters.

B LLM Prompts and Example Responses

Tables 6 to 10 illustrate all our LLM prompts along with sample inputs and responses.

[5] https://github.com/facebookresearch/BLINK
[6] https://github.com/facebookresearch/GENRE
[7] https://github.com/amazon-science/ReFinED
[8] MSNBC, AQUAINT, ACE2004, WNED-CWEB, WNED-WIKI, and the train, dev, test splits of AIDA-YAGO2

Prompt Consider the following text. Text: Nearly 17 months after he first issued his call for a "fresh start after a season of cynicism", Gov. George W. Bush ended his quest for the presidency Monday on a nearly identical note, pledging to purge { Washington } of what he cast as a crippling discord. The Texas governor claimed that Gore's decades of experience in Washington had estranged him from the rest of the country by making him too trusting of federal government and too fond of federal spending. "My opponent vows to carry his home state", Bush said. " He may win Washington, D.C., but he's not going to win Tennessee. "He forgot his roots", Bush added. "He forgot where he's from. He trusts Washington. We trust the people." Please provide me more descriptive information about { Washington } from the text above. Make sure to include Washington in your description. Answer:
Response Washington is the capital of the United States and the seat of the federal government. It is located on the Potomac River, between Maryland and Virginia. It is home to numerous monuments, memorials, and government buildings, including the White House, the Capitol Building, and the Supreme Court.

Table 6: Our zero-shot prompt with an example input and response. The mention and context inputs are from the MSNBC dataset, and the response is generated by GPT. Our zero-shot prompt is used to generate exemplars for our three-shot prompt.

Prompt

Example 1. Consider the following text.
Text: Nearly 17 months after he first issued his call for a "fresh start after a season of cynicism", Gov. George W. Bush ended his quest for the presidency Monday on a nearly identical note, pledging to purge { Washington } of what he cast as a crippling discord. The Texas governor claimed that Gore's decades of experience in Washington had estranged him from the rest of the country by making him too trusting of federal government and too fond of federal spending. "My opponent vows to carry his home state", Bush said. " He may win Washington, D.C., but he's not going to win Tennessee. "He forgot his roots", Bush added. "He forgot where he's from. He trusts Washington. We trust the people."
Please provide me more descriptive information about { Washington } from the text above.
Answer:
Washington is the capital of the United States and the seat of the federal government. It is located on the Potomac River, between Maryland and Virginia. It is home to numerous monuments, memorials, and government buildings, including the White House, the Capitol Building, and the Supreme Court.

Example 2. Consider the following text.
Text: O'Donnell and Trump have been feuding since he announced last month that Miss USA Tara Conner, whose title had been in jeopardy because of underage drinking, would keep her crown. Trump is the owner of the Miss Universe Organization, which includes Miss USA and Miss Teen USA. The 44-year-old outspoken moderator of "The View", who joined the show in September, said Trump's news conference with { Conner } had annoyed her "on a multitude of levels and that the twice-divorced real estate mogul had no right to be "the moral compass for 20-year-olds in America". Trump fired back, calling O'Donnell a "loser" and a "bully", among other insults, in various media interviews.
Please provide me more descriptive information about Conner from the text above.
Answer:
Conner is the Miss USA titleholder whose title was in jeopardy due to underage drinking. She was saved from losing her crown by Donald Trump, the owner of the Miss Universe Organization, which includes Miss USA and Miss Teen USA. Tara Conner was given a second chance by Trump and was allowed to keep her crown.

Example 3. Consider the following text.
Text: Scottish Labour Party narrowly backs referendum. STIRLING, Scotland 1996-08-31 British Labour Party leader Tony Blair won a narrow victory on Saturday when the party's Scottish executive voted 21-18 in favour of his plans for a referendum on a separate parliament for Scotland. Blair once pledged to set up a Scottish parliament if the Labour won the next general election, which must be held by May 1997. Prime Minister John Major says the 300-year-old union of the Scottish and English parliaments will be a main plank in his Conservative Party's election platform. Conservatives have only 10 of the 72 Scottish seats in parliament and consistently run third in opinion polls in Scotland behind { Labour } and the independence-seeking Scottish National Party.
Please provide me more descriptive information about { Labour } from the text above.
Answer:
The Labour Party is a centre-left political party in the United Kingdom. It is the main opposition party to the Conservative Party and is led by Tony Blair. The Labour Party has traditionally been strong in Scotland, and the Scottish Labour Party is a branch of the UK Labour Party. In the text, the Scottish Labour Party narrowly voted in favour of Tony Blair's plans for a referendum on a separate parliament for Scotland.

Now consider the following text.
Text: *left context* + { *mention* } + *right context*
Please provide me more descriptive information about { *mention* } from the text above.
Answer:

Table 7: Our three-shot prompt template, used for our main experiments detailed in section 4. All three exemplars are obtained by employing our zero-shot prompt presented in Table 6 over GPT. Due to space limitation, this prompt does not provide an example input and response. In real application, *left context*, *right context*, and *mention* are filled with their corresponding input entries.

Prompt
Gives the text and mentions within the text highlighted by <MENTION> and </MENTION>. Please give which page in Wikipedia this mention is most likely to be? Please answer me directly in this form: "mention":"Wikipedia page url". Text: Having caught the popular attention and with goodwill at a high-point , Nelsonic was able to obtain licensing from several big-name video game companies such as Sega , Nintendo ,<MENTION> Midway Games </MENTION>, and Mylstar Electronics . Answer: "Midway Games": "https://en.wikipedia.org/wiki/Midway_Games" Text: State Highway 110 or SH 110 is a state highway in the U.S. state of Texas that runs from Grand Saline to Rusk . SH 110 begins at an intersection with and in downtown Rusk and leaves the courthouse square north with US 84 , crossing on its way to a split on the northeast side of Rusk where US 84 goes off east and SH 110 turns north , out of town . The road passes <MENTION> Ponta </MENTION> and New Summerfield before crossing the county line into Smith County as it enters Troup . After a brief downtown multiplex with SH 135 , SH 110 leaves Troup going northwest through Whitehouse on its way to Tyler . Answer: "Ponta": "https://en.wikipedia.org/wiki/Ponta,_Texas" Text: Messier 49 (also known as M 49 or NGC 4472) is an elliptical galaxy located about away in the equatorial <MENTION> constellation </MENTION> of Virgo . This galaxy was discovered by French astronomer Charles Messier on February 19 , 1771 . Answer: "constellation": "https://en.wikipedia.org/wiki/Constellation" Text: <MENTION> Xinhua News Agency </MENTION> , Shanghai , April 3rd , by reporter Jierong Zhou Recently , HSBC has moved its Shanghai branch to the China Shipping Mansion in the Pudong Lujiazui financial trading district , becoming the third foreign capital bank to be approved to operate RMB business and shift to Pudong. ... Answer:
Response
"Xinhua News Agency": "https://en.wikipedia.org/wiki/Xinhua_News_Agency"

Table 8: An example of prompt and response for LLM direct execution of EL tasks, from the ACE04 dataset, and responses generated by gpt-3.5-turbo-instruct. The prompt includes an instruction, three demonstrations, and a question to be addressed. Due to space limitations, we have omitted the text in the question part during presentation.

Prompt Instruction: Gives the text and mentions within the text highlighted by <MENTION> and </MENTION>. Please select from the options below which Wikipedia page this mention is most likely to be from? Please answer me directly in this form: "(letter): Wikipedia entity name and url." And I also want you to give an explanation in the next line. Options: (1): ['Time Warner Interactive', 'https://en.wikipedia.org/wiki?curid=12642915'] (2): ['TT Games', 'https://en.wikipedia.org/wiki?curid=49108324'] (3): ['Atari Games', 'https://en.wikipedia.org/wiki?curid=304833'] (4): ['Midway Games', 'https://en.wikipedia.org/wiki?curid=430266'] (5): ['Vivendi Games', 'https://en.wikipedia.org/wiki?curid=6573837'] Text: Having caught the popular attention and with goodwill at a high-point , Nelsonic was able to obtain licensing from several big-name video game companies such as Sega , Nintendo ,<MENTION> Midway Games </MENTION>, and Mylstar Electronics . Answer: (4): ['Midway Games', 'https://en.wikipedia.org/wiki?curid=430266'] Explanation: The mention "<MENTION> Midway Games </MENTION>" in the provided text is most likely from the Wikipedia page for Midway Games. Midway Games is mentioned in the text as one of the big-name video game companies from which Nelsonic obtained licensing. The description of Midway Games in option (4) matches the context in the text, making it the most likely source. Options: (1): ['İhlas News Agency', 'https://en.wikipedia.org/wiki?curid=51484970'] (2): ['Kuwait News Agency', 'https://en.wikipedia.org/wiki?curid=3824465'] (3): ['Vietnam News Agency', 'https://en.wikipedia.org/wiki?curid=11835464'] (4): ['Hai Xia', 'https://en.wikipedia.org/wiki?curid=41850118'] (5): ['Saudi Press Agency', 'https://en.wikipedia.org/wiki?curid=2902711'] ... (99): ['BNO News', 'https://en.wikipedia.org/wiki?curid=13403728'] (100): ['Beat Holdings', 'https://en.wikipedia.org/wiki?curid=3536753'] Text: <MENTION> Xinhua News Agency </MENTION> , Shanghai , April 3rd , by reporter Jierong Zhou Recently , HSBC has moved its Shanghai branch to the China Shipping Mansion in the Pudong Lujiazui financial trading district , becoming the third foreign capital bank to be approved to operate RMB business and shift to Pudong. ... Answer:
Response (53): ['Xinhua News Agency', 'https://en.wikipedia.org/wiki?curid=263168'] Explanation: The mention "<MENTION> Xinhua News Agency </MENTION>" in the provided text is most likely from the Wikipedia page for Xinhua News Agency. The text mentions Xinhua News Agency as the source of the news report, and the description of Xinhua News Agency in option (53) matches the context in the text, making it the most likely source.

Table 9: An example of prompt and response for LLM to re-rank the returned retrieval results at Re-rank 100, from the ACE04 dataset, and responses generated by gpt-3.5-turbo-instruct. This prompt contains an instruction, a demonstration, and a question. For simplicity in presentation, we have omitted some options and text and replaced them with ellipses.

Prompt
Gives the text and mentions within the text highlighted by <MENTION> and </MENTION>. Please select from the options below which Wikipedia page this mention is most likely to be from? Please answer me directly in this form: "(letter): Wikipedia entity name and url".And I also want you to give an explanation in the next line.

Options:
(a): ['TT Games', 'https://en.wikipedia.org/wiki?curid=49108324', 'TT Games Limited is a British holding company and a subsidiary of Warner Bros. Games. ...']
(b): ['Atari Games', 'https://en.wikipedia.org/wiki?curid=304833', 'Atari Games Corporation, known as Midway Games West Inc. after 1999, was an American producer of arcade games. ...']
(c): ['Midway Games', 'https://en.wikipedia.org/wiki?curid=430266', 'Midway Games Inc., known previously as Midway Manufacturing and Bally Midway, and commonly known as simply Midway, was an American video game developer and publisher. ...']
Text: Having caught the popular attention and with goodwill at a high-point , Nelsonic was able to obtain licensing from several big-name video game companies such as Sega , Nintendo ,<MENTION> Midway Games </MENTION>, and Mylstar Electronics .
Answer: (c): ['Midway Games', 'https://en.wikipedia.org/wiki?curid=430266']
Explanation: For mention of "<MENTION> Midway Games </MENTION>", the most similar option is option (c) Midway Games. Additionally, the description in option (c) of Midway Games as an American video game developer and publisher matches the context in the text, making it the most likely source.

Options:
(a): ['China Xinhua News Network Corporation','https://en.wikipedia.org/wiki?curid=25626750', 'The China Xinhua News Network Corporation (CNC, Chinese: 中国新华新闻电视网; pinyin: zhōng guó xīn huá xīn wén diàn shì wǎng) is a state-owned television network subsidiary of Xinhua News Agency.']
...
(j): ['News agency','https://en.wikipedia.org/wiki?curid=47917', 'A news agency is an organization that gathers news reports and sells them in to subscribing news organizations, such as newspapers, magazines and radio and television broadcasters. News agencies are known for their press releases. A news agency may also be referred to as a wire service, newswire, or news service.']
Text: <MENTION> Xinhua News Agency </MENTION> , Shanghai , April 3rd , by reporter Jierong Zhou Recently , HSBC has moved its Shanghai branch to the China Shipping Mansion in the Pudong Lujiazui financial trading district , becoming the third foreign capital bank to be approved to operate RMB business and shift to Pudong . ...
Answer:

Response
(g): ['Xinhua News Agency','https://en.wikipedia.org/wiki?curid=263168']
Explanation: For mention of "<MENTION> Xinhua News Agency </MENTION>", the most similar option is option (g) Xinhua News Agency. Additionally, the description in option (g) of Xinhua News Agency as the official state news agency of the People's Republic of China matches the context in the text, making it the most likely source.

Table 10: The prompt for LLM to re-rank the returned retrieval results at Re-rank 10, from the ACE04 dataset, and responses generated by gpt-3.5-turbo-instruct. This prompt contains an instruction, a demonstration, and a question. For simplicity in presentation, we have omitted some abstracts, options and text and replaced them with ellipses.