ChatEL: Entity Linking with Chatbots

Yifan Ding, Qingkai Zeng, Tim Weninger

Department of Computer Science & Engineering
University of Notre Dame
Notre Dame, IN, USA
{yding4, qzeng, tweninge}@nd.edu

Abstract

Entity Linking (EL) is an essential and challenging task in natural language processing that seeks to link some text representing an entity within a document or sentence with its corresponding entry in a dictionary or knowledge base. Most existing approaches focus on creating elaborate contextual models that look for clues the words surrounding the entity-text to help solve the linking problem. Although these fine-tuned language models tend to work, they can be unwieldy, difficult to train, and do not transfer well to other domains. Fortunately, Large Language Models (LLMs) like GPT provide a highly-advanced solution to the problems inherent in EL models, but simply naive prompts to LLMs do not work well. In the present work, we define ChatEL, which is a three-step framework to prompt LLMs to return accurate results. Overall the ChatEL framework improves the average F1 performance across 10 datasets by more than 2%. Finally, a thorough error analysis shows many instances with the ground truth labels were actually incorrect, and the labels predicted by ChatEL were actually correct. This indicates that the quantitative results presented in this paper may be a conservative estimate of the actual performance. All data and code are available as an open-source package on GitHub at https://github.com/yifding/In_Context_EL.

Keywords: Information Extraction, Natural Language Processing, Generative Model

1. Introduction

The introduction of Large Language Models (LLMs) has injected enormous excitement and anxiety into the world of natural language processing (NLP) and artificial intelligence (AI) generally. Although their text-generation and unstructured reasoning capabilities appear to be outstanding (Raffel et al., 2020; Radford et al., 2018, 2019; Brown et al., 2020; Touvron et al., 2023), their ability to produce structured output remains underdeveloped and relatively unexplored (Yu et al., 2022; Sun et al., 2023). The entity disambiguation task of Information Extraction (IE) seeks to link text fragments that represent some real-world entity with a structured list of that entity in, say, a knowledge base or dictionary. Once linked, the existing data and its relationships within the knowledge base could be used to assist in a number of downstream processes. This is the goal of the present work: to use LLMs to link the text fragments in the documents to structured knowledge.

The merging of deep neural network models like the transformers used in LLM with the symbolic reasoning capabilities of extant knowledge bases has been dubbed by the United States Defense Advanced Research Projects Agency (DARPA) as the *Third Wave of AI* (DARPA, 2018) and as necessary for the advancement of problem solving and reasoning in AI systems (Brooks, 1981; Hitzler and Sarker, 2022). For such a vision to become reality, it is critical that LLMs be linked to the symbolic entities that they reference.

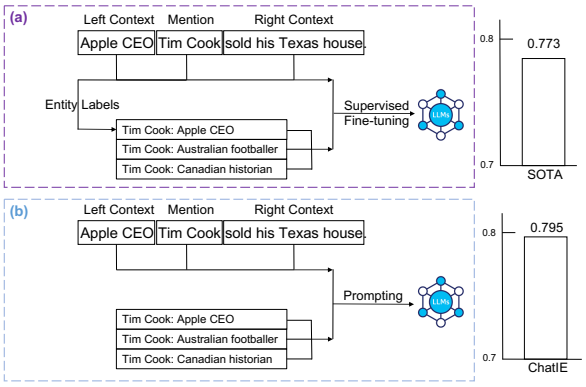

Figure 1: (a) General pipeline for supervised information extraction systems. These systems require careful modeling of the mention-text and its context and are fine-tuned on a large language model (LLM). (b) ChatEL relinquishes the context modeling entirely to the LLM and, instead, directly prompts LLM with the mention, context and entity candidates. ChatEL obtains a mean F1 score of 0.795 over ten datasets compared to 0.773 from the previous SOTA model.

Perhaps unexpectedly, citation hallucinations (Emsley, 2023) within LLMs is a related problem. In these instances, the LLM generates a fake quote or citation from a fake source. Late-breaking work in citation-generation (Mohammad et al., 2009) and text-grounding (Mun et al., 2020) are working towards reconciling generated text with structured data. Likewise, recent work in prompt engineering and orchestration

LREC-COLING 2024, pages 3086–3097

frameworks like LangChain (Chase, 2022) and Haystack (Pietsch et al., 2019) have begun to make inroads into generating structured output, like tables and lists. But in both cases, these systems do not perform information extraction because they do not link text snippets with external structured data. Recent research efforts (Qin et al., 2023; González-Gallardo et al., 2023) have utilized large generative models, specifically GPT-3.5 on information extracting tasks. However, these efforts are mainly focused on name entity recognition with tens of entity types, a far cry from the actual number (6 million) for general Wikipedia entities.

In the present work, we focus on the entity disambiguation task, one of the most difficult subtasks in IE, which requires ground text into actual entities. The state of the art in entity disambiguation is strongly rooted in supervised learning, where entity labels (Wikipedia pages or DBpedia entries) are predicted from a model trained on Wikipedia links (Ganea and Hofmann, 2017), Web hyperlinks (Wu et al., 2020), or info-boxes (Ayoola et al., 2022; Bhargav et al., 2022), considering inner-structure among entities (Hu et al., 2020). And this has been shown to work in closed-world cases where the data is clean and fully available (Hoffart et al., 2011) or in open-world scenarios where the data may be missing, but it is well-described (Logeswaran et al., 2019; Wu et al., 2020).

Another paradigm to deal with entity disambiguation task is through text generation. One of the earliest generative models to be used in entity disambiguation is called GENRE (Cao et al., 2021). The GENRE proposed a sequence-to-sequence framework that could generate an entity-label sequence from a mention-sequence conditioned with some special indicators. However, like most existing entity disambiguation work, GENRE required full training from scratch, which required an enormous amount of data and hardware resources.

However, supervised methods fail when the data is noisy, incomplete, poorly described, or rare. Recent work has found that at least 5% of the ground truth labels on the entity-recognition task of the CONLL03 dataset are incorrect (Wang et al., 2019; Zhou and Chen, 2021; Zeng et al., 2021). Likewise, at least 10% of the ground truth labels on the entity disambiguation task of the CONLL03 dataset are likely incorrect (Botzer et al., 2021; Ding et al., 2022b).

LLMs for Information Extraction Large language models like GPT-3.5 (Ouyang et al., 2022), GPT-4 (OpenAI, 2023) and others have quickly supplanted individual modeling efforts that have permeated the NLP community for years. The incredible performance of large language models in zero-shot and few-shot settings has quickly replaced the previous pre-trained and fine-tuning approaches. How to design prompts for interaction with LLMs has become a highly regarded issue. The widespread availability of these prompt-based paradigms and their profound capabilities make them a natural ally in the improvement of many information extraction tasks. Recent efforts aim to prompt large pre-trained language models for various information extraction tasks, including named entity recognition and machine reading comprehension. The idea is to create some pre-defined template to query LLMs to obtain the desired output. Type-oriented prompts (Chen et al., 2023; Sun et al., 2023) aims to find corresponding mentions for desired class while span-oriented prompts (Wang et al., 2023) aim to obtain the corresponding class for each span. Recently, PromptNER (Shen et al., 2023) combined type-oriented methods and span-oriented methods to formulate NER into a finite set of "ENTITY is TYPE", and employed a dynamic template filling to assign the corresponding relationships.

Unfortunately, adapting these frameworks to the entity disambiguation task requires some careful thought and experimentation. Entity disambiguation tasks are difficult primarily due to the unwieldy and generally undefined size of the class set. Specifically, encoding and discriminating millions of classes in entity disambiguation tasks with existing prompt methods are difficult and even unfeasible.

To solve this problem, we describe a straightforward and effective framework to utilize LLMs to assist in entity disambiguation task in this work. We propose a three-step framework based on prompting LLMs called ChatEL. ChatEL first generates a set of entity candidates for the mentions in the document. Then, it utilizes the power of the LLM to generate auxiliary content to support the selection of the corresponding entity from the candidates set.

We compared several state-of-the-art entity disambiguation models and evaluated them on ten public benchmarks. We show that ChatEL achieves comparable performance to the supervised models even without any training/fine-tuning on human-annotated data. We also conducted a further error analysis and showed two findings: 1) ChatEL was actually (sometimes) more reasonable than the answers provided on the ground truth; 2) In some cases, even ChatEL can not offer the correct prediction, the predictions of ChatEL are also highly-related to ground truth (*e.g.*, hypernym of the ground truth entity). In short, our contributions can be summarized as follows:

- We present ChatEL, a three-step Information Extraction tool that uses three-step prompts on LLMs. It successfully works on entity dis-

ambiguation task with millions of classes.

- We provide a comprehensive evaluation using ten datasets and compare the results with several state-of-the-art supervised models. Without supervised fine-tuning step, ChatEL matches and even outperforms supervised models with fine-tuning.

- We conduct a thorough error analysis. We show that the errant cases are oftentimes (arguably) more-correct than the ground truth. Also, when making mistakes, ChatEL still predicts quite close guesses on the ground truth.

2. Related Work

2.1. Entity Disambiguation

Entity disambiguation is one of the most challenging tasks in information extraction, as it aims to map annotated segments of a document to specific entities in a knowledge base. Existing research is primarily divided into two categories: improving feature extraction and refining task formulations. Key features include the interaction between mentions and context (Ganea and Hofmann, 2017; Kolitsas et al., 2018), consistency between entities and entity types (Tedeschi et al., 2021),relationship between entities and knowledge base entries (Ayoola et al., 2022), and correlation between entities (Phan et al., 2019). Traditional entity disambiguation models framed the task as entity classification (Ganea and Hofmann, 2017) while more recent work approached it as machine reading comprehension (Barba et al., 2022), dense retrieve (Wu et al., 2020), question answering (Zhang et al., 2022), and sequence-to-sequence generation (Cao et al., 2021).

2.2. Prompt-based Learning for IE

One of the pivotal developments in the field of prompt learning for information extraction was the release of GPT-3 (Brown et al., 2020). GPT-3 demonstrated remarkable capabilities in understanding and responding to natural language prompts. The benefit of prompt learning is the ability to query a language model to obtain its knowledge and understanding about a given context, showing significant ability in zero-shot and few-shot learning (Wang et al., 2023; Sun et al., 2023). Existing work on prompt learning for information extraction can be primarily categorized into two paradigms: type-oriented and span-oriented. Type-oriented methods (Ding et al., 2022a) mainly locate the given class for a certain mention within the original documents, while span-oriented methods (Cui et al., 2021) enumerate all possible spans and assign corresponding class labels.

3. Problem Definition

In this section, we present the key concepts used in this paper and formally define the entity disambiguation problem: Given a document represented as a sequence of tokens $\mathcal{D}$ and a set of subsequences $M = \{m_1, \cdots, m_n\}$ (*i.e.*, mentions) containing in $\mathcal{D}$. The goal of the entity disambiguation task is to establish a mapping for each mention $m \in M$ to its corresponding entity e in the set $\mathcal{E}$ representing entities in a knowledge base (such as Wikipedia or DBpedia).

4. ChatEL: Information Extraction with Chatbots

In the present work, we bring the power of LLMs to the information extraction task, especially entity disambiguation. The proposed ChatEL formulates entity disambiguation as a 3-step conditional selection: 1) Generate and filter entity candidates for LLM from knowledge base $\mathcal{E}$; 2) Augment each mention by extracting relevant information with a prompt for LLM; 3) Combine candidates from Step 1 with context from Step 2, forming a multi-choice question for LLM.

4.1. Step 1: Entity Candidate Generation

Given a mention m in document D, we aim to find a subset of entity candidates $\mathcal{E}_c$ corresponding to m within the KB $\mathcal{E}$. We combine two strategies to generate entity candidates. First, we deploy the Prior, which is based on the statistical information of hyperlinks (Ganea and Hofmann, 2017) and which allows us to obtain a subset of entity candidates $\mathcal{E}_p$ from $\mathcal{E}$ that exhibit syntactic similarity to the mentions within the document. However, the Prior suffers from low recall because it only considers syntactic similarity. In such cases, we also employ a dense retrieval model as a backup to augment the entity candidate generation. Specifically, we select the BLINK model (Wu et al., 2020) as our retrieval model to generate extra entity candidates $\mathcal{E}_r$. The BLINK model constructs the dense retrieval model using cleaned Wikipedia hyperlinks. Therefore, the final entity candidates set from step 1 is $\mathcal{E}_c = \mathcal{E}_p \cup \mathcal{E}_r$, which includes 10 candidates.

4.2. Step 2: Augmentation by Prompting

Since candidates are syntactically similar, distinguishing them is challenging without contextual information. To augment mentions with relevant information, we ask the LLM "What does Tim Cook represent?" (as shown in Fig. 2) to generate the auxiliary content $\mathcal{A}$ of "Tim Cook" based on the given document $\mathcal{D}$. We found that generating context information in this way has the following two advantages: 1) The content generated by the LLM is based on contextual information from the given

Table 1: Data statistics for 10 experimental datasets. Number of documents (# of Docs) and number of mentions (# of Mention).

	KORE	OKE15	OKE16	REU	RSS	ACE04	MSN	WIKI	AQU	CWEB
# of Docs	50	101	173	113	357	35	20	319	50	320
# of Mention	144	536	288	650	524	257	656	6793	727	11154

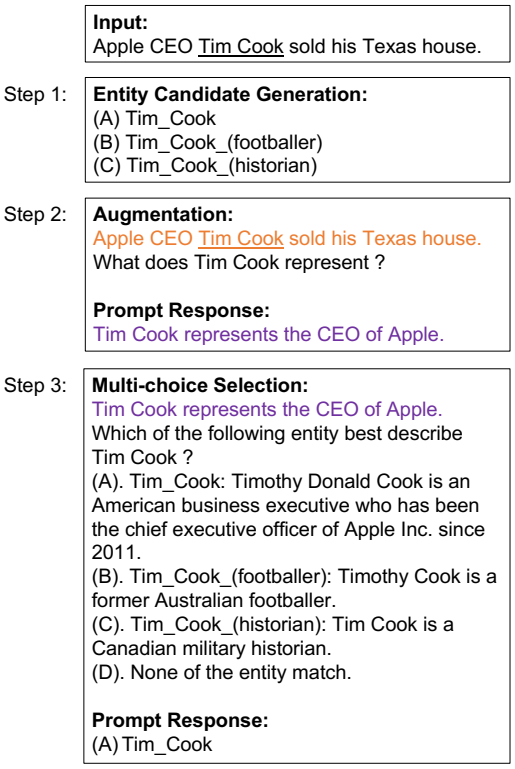

Figure 2: Pipeline of ChatEL framework: Given input document with the annotated mention, ChatEL first conducts (1) entity candidate generation step to obtain relevant entities. Then (2) an augmentation step is performed to obtain an auxiliary content of the annotated mention. Finally, (3) a multi-choice selection prompt is conducted to decide the corresponding entity of annotated mention.

document $\mathcal{D}$ and supplemented with the world knowledge encoded within the LLM.
2) The auxiliary content produced by our system is more specifically targeted towards our task. Consequently, it significantly minimizes the impact of superfluous information.

4.3. Step 3: Multiple-choice Selection by Prompting

Given entity candidates set $\mathcal{E}_c$ (from step 1) and the auxiliary content $\mathcal{A}$ (from step 2), the goal of step 3 is to select the corresponding $e \in \mathcal{E}_c$. As shown in Fig. 2, we employ an instruct-based prompt to direct the LLM to make a selection from $\mathcal{E}_c$, utilizing the auxiliary content $\mathcal{A}$. To distinguish among these entity candidates effectively, we employ the first sentence extracted from the Wikipedia page of each entity candidate as a descriptive reference. It is noted that in step 3, all entity candidates come from the subset $\mathcal{E}_c$ obtained in step 1, rather than the complete KB E. Therefore, this multi-choice setting can not address the situation where the corresponding entity e is not included in $\mathcal{E}_c$. To accommodate this situation, we include the option "None of the entity match" among the choices for handling such cases.

5. Experiments

Our proposed framework is evaluated on ten benchmarks. The experiments aim to address three research questions (RQs):
RQ1: How does the performance of ChatEL compare to baselines in entity disambiguation?
RQ2: How do the components impact the performance of ChatEL?
RQ3: What are the reasons for ChatEL failing in some cases?

5.1. Datasets

Table 1 presents the statistics of ten benchmarks we used to evaluate the entity disambiguation task. In these ten benchmarks, there are five in-domain and five out-of-domain benchmarks each. All the experiments used Wikipedia as the background Knowledge Base (KB). Following the example of Guo and Barbosa (2018), we manually removed spurious mentions that do not appear in the KB, as well as repeated documents, and empty documents without any mentions. The detail of these benchmarks are as follows:

- **MSNBC (MSN), AQUAINT(AQU), ACE2004(ACE04):** All these benchmarks have annotated from news, aiming to link entities in the news with a KB (Cucerzan, 2007; Milne and Witten, 2008; Ratinov et al., 2011).
- **WNED-WIKI (WIKI), WNED-CWEB (CWEB):** These two large auto-extracted EL evaluation sets, were developed from ClueWeb and Wikipedia by (Guo and Barbosa, 2018; Gabrilovich et al., 2013).
- **OKE-2015 (OKE15), OKE-2016 (OKE16):** They are from the Open Knowledge Extraction competition and are customized for ontology completion on DBpedia (Nuzzolese et al., 2015, 2016).

- **N3-Reuters-128 (REU), N3-RSS-500 (RSS):** N3-RSS-500 uses RSS feeds from global newspapers, covering various domains. N3-Reuters-128 includes economic news from Reuters-21587. Both datasets are manually annotated by (Röder et al., 2014).
- **KORE50 (KORE):** It contains brief, domain-varied documents from microblogging platforms like Twitter with ambiguous entity mentions (Hoffart et al., 2012).

5.2. Baselines

We compared the performance of ChatEL with the following baseline methods on entity disambiguation subtask:

- **Prior**: It is a baseline string-matching algorithm that collects the entities corresponding to a given mention by looking through the entire Wikipedia corpus. We followed preprocessing in (Ganea and Hofmann, 2017). The most frequent entity is selected as the target answer.
- **REL (van Hulst et al., 2020)**: It is an entity disambiguation and entity linking package which combines the NER method FLAIR (Akbik et al., 2019).
- **End2End (Kolitsas et al., 2018)**: It is a global entity disambiguation and entity linking system based on deep-ed (Ganea and Hofmann, 2017), but with a more robust RNN architecture that considers mention-context and mention-mention interactions.
- **GENRE (Cao et al., 2021)**: This generation-based model considers entity disambiguation task as an entity name generation process.
- **ReFinED (Ayoola et al., 2022)**: ReFinED is a recent system made by Amazon that considers Wikipedia entries as extra features to the entity disambiguation and entity linking tasks.

Since the performance of entity disambiguation highly relies on preprocessing strategies. In this work, we keep the original string from the dataset and consider all the non-empty entity as in-KB instance. During evaluation stage, we directly compare prediction entity string with the original entity string.

All data and code are available as an open-source package on GitHub at `https://github.com/yifding/In_Context_EL`.

5.3. Evaluation Metrics

To maintain a fair comparison across datasets, we use the in-KB micro-F1 score as our evaluation metric following the example of Guo and Barbosa 2018. Specifically, being *in-KB* requires that ground truth mentions correspond to existing KB entries. Empty or invalid mentions are removed in the evaluation process. Micro-F1 score is as averaged per-mention. Although a model may predict non-entities, each mention will always have some corresponding ground truth entity.

5.4. Main Results (RQ1)

The results of performance comparison over ten entity disambiguation benchmarks are presented in Table 2. The Gold-F1 reported in Table 2 shows the upper bound performance of ChatEL if the corresponding entity is included in the entity candidates set generated by step 1. The overall performance is around 90% indicating that the candidates set generated by step 1 can cover most cases in all benchmarks. We have three main observations. First, we find that the ChatEL framework using GPT-4 outperformed the second-best method by an average micro-F_1 score of $+2.2\%$. Specifically, the ChatEL obtained the best performance on three out-of-domain datasets including KORE50, Reuters-128 and RSS-500 with absolute improvements in the F1 score of 16.9%, 9.2%, and 6.6%, respectively, demonstrating the effectiveness of our methods.

Second, we can observe that ChatEL performs better on the out-of-domain benchmarks than the in-domain benchmarks while REL and End2End show strong performances on in-domain benchmarks. The main reason for this is the backbone model (word2vec) used in REL and End2End is trained on the domain-related corpus such as Wikipedia corpus. We also observe that REL and End2End have a significant drop in the performance rankings of REL and End2end on the two out-of-domain benchmarks (REU and RSS).

Third, we can notice that ChatEL is the only method without supervised fine-tuning (SFT). This indicates that ChatEL is free from human-annotated data. Compared to the baseline methods that rely on SFT, while ChatEL may not outperform them on certain benchmarks, it demonstrates greater adaptability to different domains than SFT-based baselines.

5.5. Ablation Study (RQ2)

To better understand the effects of different components, we conduct an ablation study across various aspects of the ChatEL. Due to budget constraints, we removed the two very large datasets WIKI and CWEB from the ablation study.

5.5.1. Backbone LLMs for ChatEL

We applied ChatEL on four backbone LLMs including GPT-3.5 (Ouyang et al., 2022), GPT-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2022), and LlaMa-2 (Touvron et al., 2023). We have

Table 2: Test micro-F1 scores on 10 benchmarks. Best scores are highlighted in bold, second best scores are underlined. Gold-F1 is the upper bound performance of ChatEL, as all errors stem from ground truth entities not being included in the entity candidates set generated in step 1. All experiments are re-computed to compare entity names for evaluation[1]

Model	Out-of-domain					In-domain					
	KORE	OKE15	OKE16	REU	RSS	ACE04	MSN	WIKI	AQU	CWEB	AVG
Prior	0.569	0.723	0.753	0.632	0.756	0.863	0.903	0.710	0.864	0.763	0.754
REL	0.618	0.705	0.749	0.662	0.680	**0.897**	**0.930**	0.783	**0.881**	**0.771**	0.768
End2End	0.569	0.767	0.783	0.677	0.720	0.880	0.920	0.740	0.880	0.760	0.770
GENRE	0.542	0.640	0.708	0.697	0.708	0.848	0.780	0.823	0.849	0.659	0.725
ReFinED	0.567	**0.781**	**0.794**	0.680	0.708	0.864	0.891	**0.841**	0.861	0.738	0.773
ChatIE	**0.787**	0.758	0.752	**0.789**	**0.822**	0.893	0.881	0.791	0.767	0.709	**0.795**
Gold-F1	0.880	0.903	0.903	0.911	0.921	0.969	0.970	0.944	0.981	0.943	0.932

[1]These performance reports may be different from the originally reported performance because of changes to the underlying datasets. Models tuned to out-of-date versions of the dataset may also have the names of the entries changed or removed resulting in performance degradation.

Table 3: Ablation study on 8 benchmarks (3 in-domain and 5 out-of-domain) with different backbone LLMs. The best scores are highlighted in bold, second best scores are underlined.

Backbone	Out-of-domain					In-domain			
	KORE	OKE15	OKE16	REU	RSS	ACE04	MSN	AQU	AVG
PaLM	0.728	0.662	0.665	0.742	0.767	0.852	0.814	0.685	0.739
Llama-2-70B	0.647	0.617	0.585	0.649	0.734	0.746	0.741	0.635	0.669
GPT-3.5	0.716	**0.767**	**0.770**	0.785	0.808	**0.918**	0.867	**0.791**	0.803
GPT-4	**0.787**	0.758	0.752	**0.789**	**0.822**	0.893	**0.881**	0.767	**0.806**

several observations as follows. First, both GPT-3.5 and GPT-4 consistently achieve top-tier performance across all benchmarks, ranking first and second, respectively. This indicates that language models with more parameters can significantly enhance the performance of ChatEL. Second, we found that GPT-3.5 performed similarly to GPT-4 backbone with only a small (0.3%) decrease in the average F_1 score. Thus, ChatEL has the opportunity to attain equivalent performance at a reduced expense. Last, three out-of-domain datasets (*i.e.*, KORE-50, Reuters-128, and RSS-500) found the best performance improvement with GPT-4 over GPT-3.5 ($+7.1\%$, $+0.4\%$, and $+1.4\%$ respectively). This observation means that GPT-4 has a stronger domain adaptation capability compared to GPT-3.5.

5.5.2. Effectiveness of Step 1 and Step 2

We conducted an ablation study on the eight benchmarks to verify the effectiveness of the entity candidates generation strategy (step 1) and augmentation by auxiliary content step (step 2). For step 1, we create variants of ChatEL by removing the entity candidates generated by BLINK. We also tested selecting the corresponding entity without auxiliary content for the LLM.

As shown in Table 4, in experiments in all benchmarks, removing the auxiliary content harms the performance of ChatEL. This proves that auxiliary content enhances the connections between the mention and the target entity. Note that removing the entity candidates generated by BLINK hurts the performance of ChatEL on six benchmarks. That indicates that BLINK can improve the coverage of the entity candidates set. We also have observed that ChatEL performs better without BLINK candidates on OKE15 and AQU. This is because BLINK may introduce noise into the entity candidate set, which ultimately hinders ChatEL's performance.

6. Arguing with the Teacher (RQ3)

Most studies on IE and NLP tasks mainly use quantitative analysis for model performance evaluation, often overlooking errors in ground truth (Wang et al., 2019; Zhou and Chen, 2021).. Recent analyses show a minimum of 5% error rates in benchmarks like AIDA-CONLL (Ding et al., 2022b; Botzer et al., 2021). Such error-prone datasets are commonly used in entity disambiguation and linking research, questioning whether mismatches in results are actual errors from the model or issues stemming from the dataset's own inaccuracies.

Table 4: Ablation study on 8 datasets (3 in-domain and 5 out-domain) with GPT-4 backbone. The best scores are highlighted in bold.

Ablation	Out-of-domain					In-domain			
	KORE	OKE15	OKE16	REU	RSS	ACE04	MSN	AQU	AVG
ChatIE w/o Aug. (step 2)	0.707	0.696	0.687	0.688	0.767	0.853	0.821	0.753	0.747
ChatIE w/o BLINK (step 1)	0.722	**0.769**	0.748	0.676	0.794	0.890	0.878	**0.865**	0.793
ChatIE	**0.787**	0.758	**0.752**	**0.789**	**0.822**	**0.893**	**0.881**	0.767	**0.806**

Table 5: Error Analysis of ChatEL. Data represents the absolute number of errors for each dataset and the type of the error.

	Error Type	Out-of-domain					In-domain				
		KORE	OKE15	OKE16	REU	RSS	ACE04	MSN	WIKI	AQU	CWEB
FP	**Alternative Entity**	4	26	27	34	12	11	34	507	71	1318
	Fail to Reject	8	17	19	47	42	7	2	402	12	440
FN	**Miss GT**	8	62	63	34	32	9	38	535	125	1558
	Miss Candidate	23	78	76	59	35	8	37	315	15	753

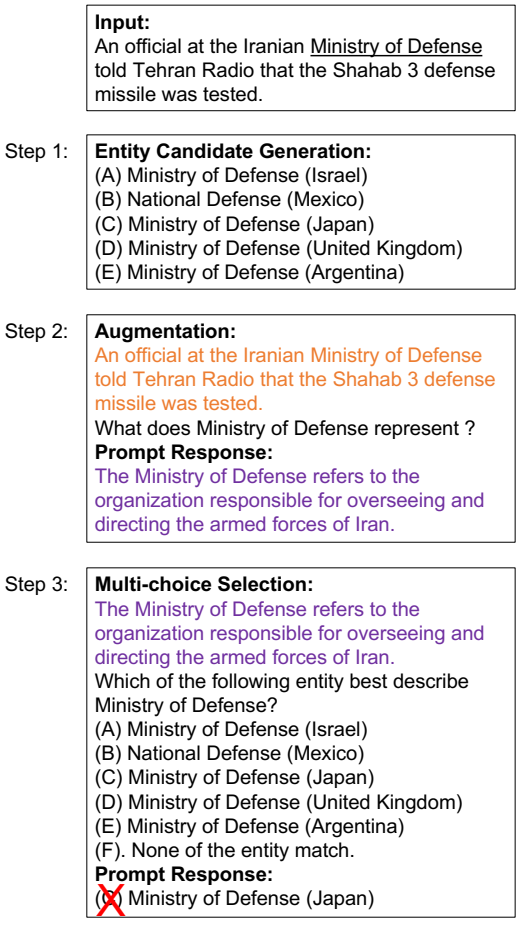

Figure 3: Error case of ChatEL predicting Ministry of Defense (Iran) vs Ministry of Defense (Japan).

6.1. Error Analysis

We first compare the model predictions with the labels in the ground truth. In a case-by-case investigation, we identified two sets of two types of errors each. In the first case, we encounter false positives where the model may miss the correct label that is present in the list of candidates (Alternative Entity) or select a label from the candidates when the ground truth label was missing and the correct answer would be to select nothing (Failure to Reject), as illustrated in Fig. 3. In the second case, we encounter false negatives where the model incorrectly predicted the absence of a label, but a label did indeed exist. In these cases, two options were possible: first the correct ground truth label could have been present from the candidate list, but the model predicted no label (Missed Ground Truth) or the correct ground truth label could have been missing from the candidate list and the model did not find it (Missed Candidate). In these four cases, the errors can be recast as conditional on the ground truth being present or not in the list of candidates from which to pick.

A detailed error analysis with this breakdown is presented in Tab. 5. From this table, we have some interesting observations. First, across all benchmarks, there does not appear to be a consistent distribution of error types. In all five out-domain datasets, the false negative error (Missed Ground Truth and Missed Candidate) appears to be more likely than the false positive error, indicating that the model hesitates to make predictions in these cases. In addition, we find most errors occur when the candidate entities do not contain the ground truth entity (row 2 and 4 in Tab. 5); that is, most errors occur when the ground truth entity is not in the list from which the label is picked—a case where the blame rests on the candidate generation step, not the model itself.

Table 6: Comparison of the ground truth label and the label predicted by the ChatEL model. We find that in many error cases the ChatEL model actually produces labels that are more accurate than the ground truth, and in several cases where the ground truth is indeed correct, the errant prediction is not far off.

Dataset	Degree of Error	Ground Truth	Prediction	Reason
ACE04	high	Ministry of Defense (Iran)	Ministry of Defense (Japan)	Step 3
ACE04	low	President of Egypt	President	Step 3
ACE04	low	Gaza City	Gaza Strip	Step 2
ACE04	low	Volvo	Volvo Cars	Step 3
KORE	low	First Ladies of Argentina	First Lady	Step 2
KORE	high	Justin Bieber	Justin I	Step 2
KORE	high	Lady Gaga	Gwen Stefani	Step 3
KORE	high	Paul Allen	NULL	Step 2
AQU	high	Cancer	Lung Cancer	Step 3
AQU	high	Tissue (biology)	Facial tissue	Step 3
CWEB	high	Head	Head (company)	Step 2
CWEB	none	Hillsborough County, Florida	Hillsborough, North Carolina	GT is incorrect
CWEB	low	Lake Wylie	Lake Wylie, South Carolina	Step 3
CWEB	none	Australia Cricket Team	Australia	GT is incorrect
MSN	none	New York City	New York	GT is incorrect
MSN	none	University of Alabama	Alabama Crimson Tide football	GT is incorrect
MSN	high	World Trade Center	Collapse of the World Trade Center	Step 3
OKE15	none	Fellow	Research Fellow	GT is incorrect
OKE15	low	Cambridge	University of Cambridge	Step 2
OKE15	none	Principal (academia)	Head teacher	GT is incorrect
OKE15	low	Faculty (academic staff)	Professor	Step 2
OKE15	none	Officer	Officer (armed forces)	GT is incorrect
OKE16	none	Director (business)	Executive director	GT is incorrect
OKE16	none	Germany	Nazi Germany	GT is incorrect
OKE16	none	Czechs	Czech Republic	GT is incorrect
OKE16	none	Sorbonne	University of Paris	GT is incorrect
REU	none	Georgia Power	Georgia (U.S. state)	GT is incorrect
REU	low	Lloyds Bank of Canada	Lloyds Bank	Step 3
RSS	none	Steve Jobs	Apple Inc.	GT is incorrect
RSS	low	Pro Bowl	Super Bowl	Step 3
RSS	none	Eric Kearney	Cincinnati	GT is incorrect
RSS	high	Cleveland Browns	Cleveland	Step 3

6.2. Case Study

In this section, we dive into ChatEL's error predictions to discern when and how ChatEL makes mistakes. We find that many of the predictions that are mismatched with the ground truth are actually more correct than the ground truth itself.

Table 6 includes some cases from benchmarks in which the prediction unmatched the ground truth. We have annotated all the error cases of ChatGPT on the KORE50 and ACE04 datasets. After being revised by human experts, we found that the F1 performance increased by 2.2% and 2.8% respectively. This demonstrates the impact of incorrect ground truth data on the actual evaluation.

We also analyze the case in which ChatGPT did make mistakes and found that many times it makes reasonable predictions. For example, when the ground truth is the President of Egypt, it predicts "President" with the same part of speech but a broader meaning. Furthermore, we analyzed what went wrong and discovered that step 3 makes the most mistakes. Additionally, step 2 sometimes makes mistakes by failing to provide useful auxiliary content, which leads to false negative predictions.

7. Discussion

In this work, we propose ChatEL, a three-step framework that leverages prompts to provide context that LLMs can use to link entity mentions from free text to their corresponding entries in a knowledge base. Unlike previous frameworks that produce complicated models to properly contextualize mention-text, the ChatEL framework simply replaces that complicated context model with an LLM with outstanding results. Unlike existing state-of-the-art models, ChatEL does not require any fine-tuning and is more accurate on average. Furthermore, the detailed analysis in Sec. 6 appears to indicate that quantitative results presented in the results section may be overcounting false positives due to errors in the ground truth. As a result, we believe that the performance metrics presented in the present work are a conservative estimate of the actual performance of ChatEL.

Ethics Statement

Code and analysis are publicly accessible on GitHub, ensuring reproducibility. Characterized as low-risk, it utilized publicly available datasets, curated from news and web sources, containing no personally identifiable information, and resistant to falsification or misuse for misleading/libelous info. The work primarily impacts text generative models' reliability, improving them by linking to curated knowledge bases, addressing issues like hallucinations in LLMs and enhancing fine-grained information tasks.

8. Bibliographical References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

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.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. ExtEnD: Extractive entity disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2478–2488, Dublin, Ireland. Association for Computational Linguistics.

G P Shrivatsa Bhargav, Dinesh Khandelwal, Saswati Dana, Dinesh Garg, Pavan Kapanipathi, Salim Roukos, Alexander Gray, and L Venkata Subramaniam. 2022. Zero-shot entity linking with less data. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1681–1697, Seattle, United States. Association for Computational Linguistics.

Nicholas Botzer, Yifan Ding, and Tim Weninger. 2021. Reddit entity linking dataset. *Information Processing & Management*, 58(3):102479.

Rodney A Brooks. 1981. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17(1-3):285–348.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Harrison Chase. 2022. LangChain.

Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han, and Le Sun. 2023. Learning in-context learning for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13661–13675, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

DARPA. 2018. Ai next campaign. *DARPA Website*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Haitao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2022a. Prompt-learning for fine-grained entity typing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages

6888–6901, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yifan Ding, Nicholas Botzer, and Tim Weninger. 2022b. Posthoc verification and the fallibility of the ground truth. In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, pages 23–29, Seattle, WA. Association for Computational Linguistics.

Robin Emsley. 2023. Chatgpt: these are not hallucinations–they're fabrications and falsifications. *Schizophrenia*, 9(1):52.

Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The World Wide Web Conference*, WWW '19, page 438–447, New York, NY, USA. Association for Computing Machinery.

Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. *Advances in Neural Information Processing Systems*, 35:15460–15475.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0).

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G Moreno, and Antoine Doucet. 2023. Yes but.. can chatgpt identify entities in historical documents? *arXiv preprint arXiv:2303.17322*.

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

Pascal Hitzler and Md Kamruzzaman Sarker. 2022. *Neuro-symbolic artificial intelligence: The state of the art*. IOS Press.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 545–554, New York, NY, USA. Association for Computing Machinery.

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.

Linmei Hu, Jiayu Ding, Chuan Shi, Chao Shao, and Shaohua Li. 2020. Graph neural entity disambiguation. *Knowledge-Based Systems*, 195:105620.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. 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.

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.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th*

ACM Conference on Information and Knowledge Management, CIKM '08, page 509–518, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592, Boulder, Colorado. Association for Computational Linguistics.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819.

Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. 2015. Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges*, pages 3–15, Cham. Springer International Publishing.

Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Robert Meusel, and Heiko Paulheim. 2016. The second open knowledge extraction challenge. In *Semantic Web Challenges*, pages 3–16, Cham. Springer International Publishing.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Minh C. Phan, Aixin Sun, Yi Tay, Jialong Han, and Chenliang Li. 2019. Pair-linking for collective entity disambiguation: Two could be better than all. *IEEE Transactions on Knowledge and Data Engineering*, 31(7):1383–1396.

Malte Pietsch, Timo Möller, Bogdan Kostic, Julian Risch, Massimiliano Pippi, Mayank Jobanputra, Sara Zanzottera, Silvano Cerza, Vladimir Blagojevic, Thomas Stadelmann, Tanay Soni, and Sebastian Lee. 2019. Haystack: the end-to-end NLP framework for pragmatic builders.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

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.

Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3 - a collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3529–3533, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021. Named Entity Recognition for Entity Linking: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*,

pages 2584–2596, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2197–2200, New York, NY, USA. Association for Computing Machinery.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163, Hong Kong, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2022. Jaket: Joint pre-training of knowledge graph and language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11630–11638.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Qingkai Zeng, Mengxia Yu, Wenhao Yu, Tianwen Jiang, and Meng Jiang. 2021. Validating label consistency in ner data annotation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 11–15.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022. EntQA: Entity linking as question answering. In *International Conference on Learning Representations*.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5381–5392, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.