RAED: Retrieval-Augmented Entity Description Generation for Emerging Entity Linking and Disambiguation

Karim Ghonim[1]   **Pere-Lluís Huguet Cabot[1]**   **Riccardo Orlando[1]**   **Roberto Navigli[1,2]**

Sapienza University of Rome
[1]`surname@diag.uniroma1.it`

Babelscape, Italy
[2]`surname@babelscape.com`

Abstract

Entity Linking and Entity Disambiguation systems aim to link entity mentions to their corresponding entries, typically represented by descriptions within a predefined, static knowledge base. Current models assume that these knowledge bases are complete and up-to-date, rendering them incapable of handling entities not yet included therein. However, in an ever-evolving world, new entities emerge regularly, making these static resources insufficient for practical applications. To address this limitation, we introduce RAED, a model that retrieves external knowledge to improve factual grounding in entity descriptions. Using sources such as Wikipedia, RAED effectively disambiguates entities and bases their descriptions on factual information, reducing the dependence on parametric knowledge. Our experiments show that retrieval not only enhances overall description quality metrics, but also reduces hallucinations. Moreover, despite not relying on fixed entity inventories, RAED outperforms systems that require predefined candidate sets at inference time on Entity Disambiguation. Finally, we show that descriptions generated by RAED provide useful entity representations for downstream Entity Linking models, leading to improved performance in the extremely challenging Emerging Entity Linking task.

1 Introduction

Many Natural Language Processing (NLP) tasks require access to world knowledge to perform effectively. Despite their impressive capabilities, Large Language Models (LLMs) struggle with emerging knowledge that was not present in their training data (Zaporojets et al., 2022; Gekhman et al., 2024). This limitation is particularly evident in scenarios where accurate and up-to-date information is essential, such as fact verification, knowledge base completion, and entity-centric tasks (Wang et al., 2024b; Sun et al., 2024; Scirè et al., 2024a). To

address this gap, retrieval-based methods are gaining traction, as they enable models to incorporate external, up-to-date information at inference time, improving their ability to handle the dynamic nature of knowledge (Lewis et al., 2020; Izacard and Grave, 2021; Zhang et al., 2022).

One such scenario is creating accurate entity representations, which is essential for knowledge-intensive tasks. Entity Disambiguation (ED) and Entity Linking (EL), in particular, both rely heavily on titles and definitions to link mentions in text to entities in a predefined knowledge base, typically assuming that the underlying knowledge base (e.g., Wikipedia) is both complete and accurate (Wu et al., 2020; De Cao et al., 2021; Procopio et al., 2023).

However, this assumption breaks down for emerging or poorly described entities, as well as in lower-resource languages, limiting the robustness and scalability of existing systems. For example, consider querying a system trained before the Covid-19 pandemic with the sentence: *By this point, most people have had at least one brush with Covid-19.* Since *Covid-19* is an emerging entity absent from the training data, the model fails to correctly resolve the mention.

Similarly, candidate retrieval-based approaches (Zhang et al., 2022; Wang et al., 2024a; Orlando et al., 2024) fail when they rely on outdated or incomplete entity inventories. Consider our example again: if an entry for *Covid-19* is not available in the knowledge base, the model is forced to select from incorrect candidates. Instead, if we retrieve general textual knowledge, such as *Covid-19 being shorthand for 'coronavirus disease 2019'*, and condition the model on this relevant context, we enable it to generate more accurate and informative outputs, even for previously unseen entities.

To address this issue, we propose RAED (Retrieval-Augmented Entity Description), a retrieval-based framework designed to tackle the emerging entity challenge. Given a single men-

tion of an entity, RAED generates a title and a definition by retrieving high-quality, contextually relevant information from external sources such as Wikipedia. Unlike traditional approaches, which are restricted to pre-constructed entity inventories (i.e., knowledge bases), our method focuses on generating meaningful descriptions, even for entities absent from the inventory.

We summarize our contributions as follows:

- **Retrieval-Augmented Entity Description Generation.** We introduce a retrieval-augmented framework that generates entity descriptions (titles and definitions) for emerging entities, enabling effective disambiguation even in the absence of pre-defined candidates.

- **A novel retrieval training paradigm for Entity Disambiguation.** We propose a retriever trained on in-context mentions of entities in Wikipedia as positive retrieval targets. This shifts the focus from retrieving static entity summaries to retrieving diverse, contextually relevant passages.

- **Bridging retrieval and generation for emerging entities.** Our approach reduces dependence on static knowledge bases by dynamically incorporating retrieved contextual passages into the generation process. This enables RAED to generate descriptions that can be utilized effectively for Emerging Entity Linking (EEL).

In the hope of fostering research in Entity Description Generation, we release our code and models at `https://github.com/SapienzaNLP/RAED`.

2 Related Work

2.1 Retrieval-Augmented Information Extraction

Retrieval-based architectures have significantly advanced Information Extraction (IE) tasks by incorporating external knowledge at inference time. Dense retrieval models, such as EntQA (Zhang et al., 2022) and ReLiK (Orlando et al., 2024), have demonstrated improvements in linking mentions to entities by retrieving high-quality candidates from large-scale indices. Our approach also builds on the retrieval-augmented paradigm; however, instead of retrieving target candidate entities, it leverages general-purpose passages. This enables not only

disambiguation, but also the generation of new entity titles and definitions. As a result, it offers a more flexible and scalable alternative to traditional methods that rely on predefined entity inventories. This also opens the door to automatically updating entity inventories, which frequently become outdated and are costly to maintain manually.

2.2 Generative Approaches for Entity Disambiguation

Autoregressive models have emerged as strong alternatives to classification-based ED systems. GENRE (De Cao et al., 2021) formulates ED as a sequence-to-sequence task, generating existing entity titles directly. More recently, FusionED (Wang et al., 2024a) employs an encoder-decoder model that uses candidate entity descriptions retrieved from an inventory. The encoder models interactions between the context and each candidate, producing individual representations, while the decoder selects the best match by generating its index from the ordered candidate list. Our approach builds on these insights by learning to generate entity descriptions. However, unlike FusionED, which narrows the search space by retrieving candidate entities directly, RAED enriches its input with passages retrieved from broad informative text. This enables it to handle emerging entities that are absent from the inventory, a capability current models lack.

2.3 Handling Emerging Entities

A core challenge in ED and EL is handling new entities that are missing from the knowledge base at training time, known as emerging entities. Traditional systems assume a static knowledge base, making them ineffective for newly introduced entities. To study this issue, TempEL (Zaporojets et al., 2022) introduced a temporally segmented benchmark that evaluates the ability of models to adapt to new entities over time. EDIN (Kassner et al., 2022) extended this by proposing an end-to-end framework and benchmark to discover and index unknown entities, highlighting the importance of dynamically updating entity representations. To the best of our knowledge, the EDIN benchmark has not been publicly released. Despite these contributions, existing approaches do not address the challenge of automatically creating entries for missing entities.
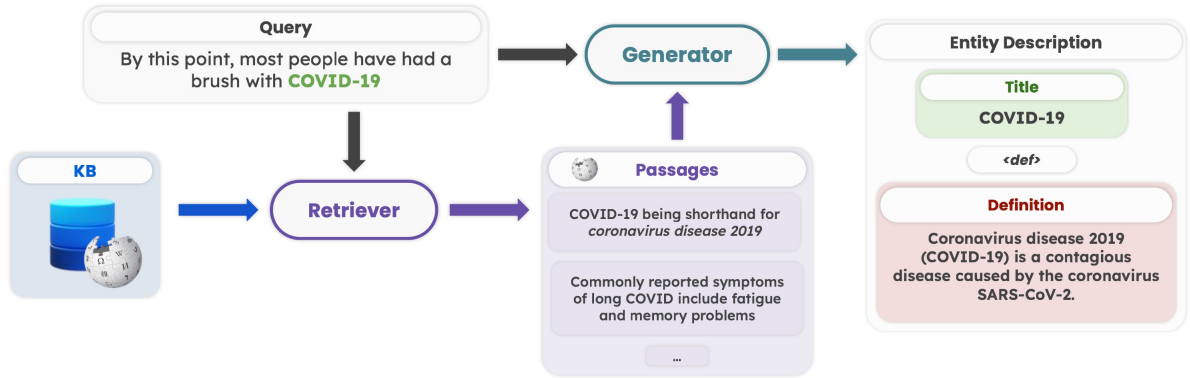
Figure 1: The Retrieval-Augmented Entity Descriptions (RAED) pipeline.

3 RAED

In this section, we introduce RAED, our model for Entity Description Generation. RAED retrieves highly informative passages from a knowledge base (i.e., Wikipedia) and provides them as additional context to the generator to ensure that:

1. Generated descriptions are grounded in factual knowledge, thus limiting hallucinations.

2. Ambiguous entities are more effectively disambiguated, particularly when the original context alone is insufficient.

3.1 Passage Index

As discussed, our goal is to construct an external index that functions as a non-parametric memory. Following prior work (Lewis et al., 2020; Izacard and Grave, 2021), we build this index using Wikipedia[1] by extracting the text from each page along with its *wikilinks*, i.e., hyperlinks to other Wikipedia pages. Previous studies have used wikilinks as pseudo-labels, treating them as disambiguation signals for linked spans (Wu et al., 2020; De Cao et al., 2021; Huguet Cabot and Navigli, 2021), while Izacard et al. (2021) leveraged entire entity pages as retrieval targets for training. In contrast, we leverage wikilinks as retrieval targets by retrieving passages where an entity is explicitly mentioned, rather than relying solely on its main Wikipedia page or description. In EL, retrieving entity descriptions is a high-precision objective (Zhang et al., 2022; Orlando et al., 2024; Wang et al., 2024a), but disambiguation fails when descriptions are missing or not retrieved. By shifting from descriptions to passages that provide additional context, i.e., either by directly mentioning

the entity or containing relevant information, we enable the generator to produce more meaningful descriptions, even for entities not present in the inventory.

To construct our index, we follow these steps:

Wikipedia Parsing We use two libraries, `wikiextractor` (Attardi, 2015) and `mwparserfromhell` (Kurtovic, 2013), to extract raw text from Wikipedia pages, including wikilinks and boundary annotations. Disambiguation pages and other non-informative content are filtered out, yielding approximately 6 million extracted pages.

Entity Filtering To ensure a manageable index size, we divide each page into non-overlapping 100-word windows and retain only those that contain entities present in our datasets, either as annotated or candidate entities. This step results in a corpus of 88 million windows.

Similarity Filtering To further refine the index and retain meaningful content, we limit the number of passages per entity to 10. Passages are ranked by their similarity to the entity's definition using a sentence similarity model,[2] and only the top-ranked ones are retained. The resulting index comprises 9 million passages.

It is important to note that while wikilinks are used during index construction, the retriever operates only on the raw text, without any direct access to the wikilinks or the associated entities.

3.2 Retriever

For the Retriever component, we follow a retrieval paradigm similar to that of DPR (Karpukhin et al.,

---

[1] English Wikipedia dump released on Jan 20, 2023.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

2020, Dense Passage Retrieval), which uses an encoder to produce dense representations of queries and passages. In our setup, the input text q contains the target entity e for which we want to retrieve relevant passages. We use special tokens, i.e., `[DEFINE]` and `[/DEFINE]`, to mark the beginning and end, respectively, of the entity mention. For example, the query *most people have had at least one brush with <u>Covid-19</u>* is represented as `most people have had at least one brush with [DEFINE] Covid-19 [/DEFINE]`.

Given $q \in \mathcal{Q}$ as our query in a collection of queries $\mathcal{Q}$ and a passage $p \in \mathcal{P}$ in a collection of passages $\mathcal{P}$ that corresponds to the Wikipedia passages or windows, the Retriever model computes:

$$E_Q(q) = \text{Retriever}(q), \ E_P(p) = \text{Retriever}(p)$$

and ranks the most relevant passages with respect to q using the similarity function $\text{sim}(q,p) = E_Q(q)^\top E_P(p)$, where the contextualized hidden representation of a query q and a passage p are the average of the encodings for the tokens in each of the two sequences computed by the same $\text{Retriever}(\cdot)$ Transformer encoder.

We train the Retriever employing a contrastive cross-entropy loss similar to SimCSE (Gao et al., 2021) as a training objective. The $\mathcal{L}_{Retriever}$ loss for q is defined as:

$$-\sum_{p^+ \in \mathcal{P}^+(q)} \log \frac{e^{\text{sim}(q,p^+)}}{e^{\text{sim}(q,p^+)} + \sum_{p^- \in P^-(q)} e^{\text{sim}(q,p^-)}} \quad (1)$$

where $\mathcal{P}^+(q)$ are the gold passages of the entity present in q, and $P^-(q)$ is the set of negative examples for q constructed using in-batch negatives from gold passages of other queries.

3.3 Entity Description Generation

We formalize the entity description task as follows: given a mention m in a query q and a set of passages $\mathcal{P}(q)$ relevant to q, RAED aims to generate a target description d for the corresponding entity:

$$P(d \mid q, \mathcal{P}(q)) = \prod_{k=1}^{|d|} P(d_k \mid d_{0:k-1}, q, \mathcal{P}(q))$$

Here, q refers to the same query introduced in Section 3.2, used as input for retrieving relevant passages. Specifically, we first retrieve the most relevant passages $\mathcal{P}(q)$, and then use those passages, along with the query q, to condition the generation

of the entity description. The target description d consists of the Wikipedia title and the article's opening sentence, separated by a special token `<def>`, as shown in Figure 1.

3.4 RAED for Entity Disambiguation

Given that our objective is to generate a meaningful description for a given entity mention, our system indirectly learns to disambiguate that entity. Hence, here we show that RAED, although trained against a generative objective, can also be applied to discriminative tasks such as ED without the need for additional training. ED aims to identify the correct entity from a set of candidates $\mathcal{E}(m)$ for a given mention m. To directly utilize the descriptions generated by RAED for ED, we employ sentence-transformers[3] to compute a similarity score between the generated description $\hat{d}$ and a candidate reference description d_e of the entity e. We first encode both the generated description and all the candidates provided for m into vector representations (i.e., embeddings). We then select the candidate with the highest cosine similarity to the generated description, as follows:

$$\hat{e} = \arg\max_{e \in \mathcal{E}(m)} \text{sim}(\hat{d}, d_e) \quad (2)$$

4 Experimental Setup

In this section, we outline the experimental setup for training and evaluating RAED. Specifically, we describe the training procedures for the retriever and generator components (Sections 4.1 and 4.2), the datasets used (Section 4.3), and the evaluation metrics used (Section 4.4).

4.1 Retriever

We train the E5[4] (Wang et al., 2022) encoder directly on our Passage Index described in Section 3.1, using a self-supervised approach similar to Gao et al. (2021). Specifically, we treat the collection of passages $\mathcal{P}$ as our set of queries $\mathcal{Q}$. For each query $q \in \mathcal{Q}$, we consider as positive passages those $p \in \mathcal{P}^+(q)$ that refer to the same entity e as q. Negative passages are sampled from passages within the same batch that do not mention the same entity. We optimize the Noise Contrastive Estimation (NCE) loss (see Eq. 1 in Section 3.2) using 400 negatives per batch. The encoder is trained for

---

[3] https://huggingface.co/Alibaba-NLP/gte-modernbert-base

[4] https://huggingface.co/intfloat/e5-base

a maximum of 100,000 steps using RAdam (Liu et al., 2020) with a learning rate of $2 \cdot 10^{-5}$ and a linear learning rate decay schedule.

4.2 Entity Description Generation

Our experiments explore various text generation models for the entity description generation task. Specifically, we evaluate both encoder-decoder models, including T5-large[5] (Raffel et al., 2020) and Flan-T5-large[6] (Chung et al., 2024), and decoder-only models, such as SmolLM2-360M[7] (Allal et al., 2025) and Llama-3.2-1B[8] (Dubey et al., 2024). We provide further details about the models used in Appendix A. For each input text in AIDA (referred to here as the query q), we combine q with its top 10 retrieved passages to create the input for RAED. We include an ablation study on the number of retrieved passages in Appendix C. We fine-tune each model on the AIDA training split to generate a description for the mention m within q for 100,000 steps with a batch size of 32 input texts, using the AdamW optimizer (Loshchilov and Hutter, 2019) and a learning rate of $1 \cdot 10^{-6}$. To assess the impact of retrieved passages, we train each model to generate descriptions using the same data and setup, once with retrieval (denoted as Model_{RAED}) and once without retrieval (denoted as Model).

Integration of Retrieved Passages For encoder-decoder models, we integrate retrieved passages using the Fusion-in-Decoder (FiD) approach (Izacard and Grave, 2021), in which each retrieved passage is concatenated with the query, independently encoded, and then passed to the decoder. For decoder-only models, we incorporate the retrieved passages directly into the prompt. An example of the prompt used is provided in Appendix A.

4.3 Datasets

Wikipedia As explained in the previous section, we leverage the entire English Wikipedia to construct our index, which is used both to train the retriever and as the source of additional context in our RAED framework.

AIDA CoNLL-YAGO We adopt the experimental setup of De Cao et al. (2021), utilizing the

standard AIDA-CoNLL splits (Hoffart et al., 2011) for training (AIDA-train), model selection (AIDA-testa), and in-domain evaluation (AIDA-testb). AIDA is a widely used benchmark for ED and EL, allowing consistent comparison with prior work and ensuring a robust evaluation framework.

Out-of-Domain (OOD) Datasets For out-of-domain ED evaluation, we use MSNBC (Cucerzan, 2007), AQUAINT (Milne and Witten, 2008), ACE2004 (Ratinov et al., 2011), WNED-CWEB (CWEB), and WNED-WIKI (WIKI), as curated by Alani et al. (2018) and Gabrilovich et al. (2013). These datasets span a diverse range of domains and text styles, providing a challenging benchmark for assessing model generalization. To ensure consistency with prior work, we use the same candidate sets originally introduced by Le and Titov (2018).

TempEL We also use TempEL (Zaporojets et al., 2022) which provides ten yearly snapshots of English Wikipedia entities, enabling the study of temporal dynamics in EL. These snapshots allow us to simulate the introduction of new entities into the knowledge base and to assess our models' ability to generate descriptions for entities not seen at training time. TempEL also provides two types of entities: *continual entities*, which appear in all TempEL temporal snapshots, and *emerging entities*, which are newly introduced in specific snapshots and serve as the focus of our experiments.

4.4 Evaluation Metrics

We evaluate the quality of the generated entity descriptions using the following metrics:

Natural Language Generation (NLG) We employ traditional metrics like BLEU and ROUGE (Papineni et al., 2002; Lin, 2004) which measure n-gram overlap, capturing lexical similarity. However, these metrics rely on exact string matches and may not fully reflect the quality of the generated text. To address this limitation, we assess the Semantic Similarity (SIM)[9] by calculating the cosine similarity between sentence embeddings of the generated and reference descriptions, focusing on semantic alignment. Additionally, we utilize BERTScore (Zhang et al., 2019), which evaluates token-level semantic similarity using contextualized embeddings from a pre-trained BERT model. By incorporating these complementary metrics, we

---

[5] https://huggingface.co/google-t5/t5-large
[6] https://huggingface.co/google/flan-t5-large
[7] https://huggingface.co/HuggingFaceTB/SmolLM2-360M
[8] https://huggingface.co/meta-llama/Llama-3.2-1B

---

[9] We use the same model in introduced in Section 3.4.

Dataset	Model	Retrieval	BL	R-1	BS	SIM	F-NLI
$AIDA_{testb}$	T5	✗	58.0	73.5	92.6	93.2	22.1
	$T5_{RAED}$	FiD	70.9	82.7	96.2	95.3	49.9
	Flan-T5	✗	60.5	77.5	93.1	92.9	24.8
	$Flan\text{-}T5_{RAED}$	FiD	70.6	82.5	96.2	95.2	50.2
	SmolLM2	✗	61.2	74.1	94.3	92.8	27.4
	$SmolLM2_{RAED}$	Prompt	72.1	83.1	95.6	94.3	52.9
	Llama-3.2	✗	63.2	76.8	95.8	93.7	33.8
	$Llama\text{-}3.2_{RAED}$	Prompt	**73.9**	**84.0**	**96.9**	**95.4**	**55.5**
OOD	T5	✗	21.0	45.8	88.1	81.2	-8.2
	$T5_{RAED}$	FiD	45.1	63,8	92,4	87.8	25.6
	Flan-T5	✗	22.9	48.1	89.5	81.9	-3.7
	$Flan\text{-}T5_{RAED}$	FiD	45.9	64.4	92.7	88.1	25.9
	SmolLM2	✗	24.4	47.0	88.7	83.0	-5.5
	$SmolLM2_{RAED}$	Prompt	42.8	61.5	97.5	85.1	21.3
	Llama-3.2	✗	28.4	51.6	90.3	84.5	-2.0
	$Llama\text{-}3.2_{RAED}$	Prompt	**45.1**	**63.3**	**92.6**	**87.5**	**27.5**

Table 1: Evaluation results of various models on the AIDA test split ($AIDA_{testb}$) and out-of-domain (OOD) datasets (MSNBC, AQUAINT, ACE2004, CWEB, and WIKI). Metrics include BLEU (BL), ROUGE-1 (R-1), BERTScore (BS), Semantic Similarity (SIM), and Factual-NLI (F-NLI). Models with ✗ do not use retrieved passages. **Bold** indicates the best performance. All reported models were fine-tuned on the AIDA training set.

aim to provide a comprehensive evaluation of the quality of the generated descriptions.

Natural Language Inference (NLI) Inspired by prior work (Chen and Eger, 2023; Scirè et al., 2024b), we evaluate the logical relationship between the generated and reference descriptions using a pre-trained NLI model.[10] Specifically, we compute the difference between the probabilities assigned to the entailment and contradiction classes to assess factual consistency. We refer to this metric as Factual-NLI (F-NLI).

InKB Micro F1 This metric, specific to Entity Disambiguation (ED), computes the micro-averaged F1 score over mentions correctly linked to entities present in the knowledge base (InKB).

5 Results

In this section, we present a three-fold evaluation of RAED. First, we assess the quality and factual accuracy of the generated entity descriptions. Next, we evaluate their effectiveness in downstream tasks, including ED and EEL. Finally, we conduct a qualitative evaluation using an LLM-as-a-Judge frame-

work to compare RAED's descriptions against descriptions generated without retrieval.

5.1 Entity Description Generation

In Table 1, we present the results of the Entity Description Generation task across various configurations of RAED. Our retrieval strategy consistently improves performance across all models and sizes, as evidenced by increases in all NLG metrics. This indicates that the descriptions generated by RAED are more lexically and semantically aligned with the gold descriptions. Notably, we observe an average increase of 27.5 points in Factual-NLI scores across RAED models, suggesting that descriptions generated by RAED are significantly more factually consistent with the reference descriptions. We attribute these improvements to RAED's grounded generation process, which reduces reliance on parametric memory, mitigates hallucinations, and leads to improved factual consistency. Similar gains are observed across all metrics for out-of-domain (OOD) datasets, highlighting RAED's capability to generate semantically and factually accurate descriptions for entities unseen during training.

Efficiency RAED provides not only performance gains but also efficiency benefits. It outperforms

Model	Retrieval	In-domain AIDA-B	Out-of-domain MSNBC	Out-of-domain AQUAINT	Out-of-domain ACE2004	Out-of-domain CWEB	Out-of-domain WIKI	AVG ALL	AVG OOD
GENRE	-	88.6	88.1	77.1	82.3	71.9	71.7	80.0	78.2
FusionED	*Candidates*	<u>91.7</u>	<u>92.4</u>	82.0	87.1	75.8	78.6	84.6	83.2
T5	-	90.1	91.3	78.4	89.4	74.4	72.2	82.6	81.1
T5$_{RAED}$	*Passages*	<u>91.7</u>	92.3	82.0	<u>90.7</u>	<u>77.3</u>	<u>78.7</u>	**85.5**	**84.2**
FLAN-T5	-	89.4	91.5	79.5	**91.4**	74.7	73.4	83.3	82.1
FLAN-T5$_{RAED}$	*Passages*	**91.8**	**93.2**	80.3	90.3	77.1	78.1	<u>85.1</u>	<u>83.8</u>
SmolLM2	-	89.1	91.6	82.3	85.8	76.3	77.5	83.8	82.7
SmolLM2$_{RAED}$	*Passages*	91.0	91.9	<u>83.6</u>	88.3	76.0	77.1	84.5	83.4
Llama-3.2	-	91.0	92.1	82.1	<u>90.7</u>	74.9	75.0	84.3	83.0
Llama-3.2$_{RAED}$	*Passages*	<u>91.7</u>	92.3	**84.3**	89.5	**78.1**	**79.0**	85.0	83.4

Table 2: InKB Micro F1 comparison of different models fine-tuned on the AIDA training set and evaluated on both in-domain and out-of-domain datasets. Results for GENRE and FusionED are reported as presented in their respective publications. **Bold** values indicate the best performance, while <u>underlined</u> values denote the second best.

both same-sized and significantly larger models that do not use retrieval. For instance, RAED, with SmolLM2 (360M parameters) as the generator and E5-base (109M parameters) as the retriever (totaling approximately 469M parameters), outperforms Llama-3.2 (1B parameters) across all metrics and achieves a 23.3-point gain in factuality, despite being less than half its size. This result demonstrates that incorporating external knowledge while reducing reliance on parametric memory can improve both performance and efficiency.

5.2 Entity Disambiguation

Table 2 presents our results using RAED as a discriminator for ED, as detailed in Section 3.4. Compared to models without retrieval, RAED variants show improvements across nearly all benchmarks, in both in-domain and out-of-domain settings. The most notable gains are observed with T5$_{RAED}$ and FLAN-T5$_{RAED}$, which outperform T5 and FLAN-T5 by 2.9 and 1.8 points, respectively.

Remarkably, RAED models not only match but also slightly outperform systems like FusionED, which are specifically trained for ED and provided with candidate entities as input. For example, although it uses the same underlying architecture (FLAN-T5-large), FLAN-T5$_{RAED}$ outperforms FusionED on most out-of-domain datasets, achieving an average improvement of 0.6 points, while also slightly surpassing it on the in-domain benchmark (91.8 vs. 91.7). This result is particularly noteworthy because FusionED is explicitly trained to

select from a set of candidate entities, whereas RAED relies solely on retrieved passages to generate descriptions and learns to disambiguate mentions implicitly, without any direct ED supervision during training. Furthermore, our approach enables models like T5$_{RAED}$ to outperform larger and more recent models such as Llama-3.2-1B, achieving a 1.2-point overall improvement and a 1.2-point gain in the out-of-domain setting.

These results, obtained without access to candidate entities at inference time, show that RAED exhibits strong generalizability and versatility. This highlights a key advantage of our approach: RAED can generate meaningful and discriminative descriptions using retrieved passages, eliminating dependence on predefined candidates even when the correct entity is absent from the knowledge base.

5.3 Emerging Entity Linking

In this section, we evaluate the effectiveness of RAED in handling emerging entities. As detailed in Section 4.3, we use the temporal splits from Wikipedia provided by the TempEL dataset. For each split, we define an entity as *emerging* if it did not appear in any of the previous snapshots (e.g., entities added between 2020 and 2021 for the 2021 split). Although TempEL includes splits from 2013 to 2022, we focus our analysis on those from 2019 onward to better simulate scenarios in which emerging entities are unseen during model pre-training, particularly for models such as T5.

We evaluate EL performance on temporal test

Model	Desc	2020	2021	2022
2019	T5	51.2	37.8	44.5
	T5$_{RAED}$	**57.0**	**47.4**	**53.5**

Table 3: We evaluate EL performance (accuracy@64) on TempEL test sets (2020–2022) using descriptions generated by T5-Large (T5) and RAED (using T5-Large). Results in **bold** indicate the best performance for generated descriptions.

Model	Win Rate
T5$_{RAED}$	40%
Wikipedia (Gold)	60%
T5	9%
Wikipedia (Gold)	91%
T5$_{RAED}$	82%
T5	18%

Table 4: Win rates (%) from LLM-as-a-Judge experiments comparing RAED, T5 without retrieval, and Gold Wikipedia descriptions.

sets using the TempEL model and two types of descriptions: (1) descriptions generated by RAED, and (2) descriptions generated by models trained without retrieval. The models used are the same as those reported in Table 1. Following Zaporojets et al. (2022), we report accuracy@64 as our evaluation metric, which measures whether the target entity appears within the top-64 retrieved candidates. Additional details can be found in Appendix D. To ensure a fair evaluation and prevent data leakage, we exclude from the index any passages taken from the Wikipedia pages of the emerging entities.

Table 3 shows that RAED significantly outperforms the model with no access to retrieved knowledge, achieving an average improvement of 8.1 points. This shows that by leveraging retrieval, RAED generates descriptions with stronger semantic grounding, which in turn improves disambiguation and linking. This makes it particularly effective for the challenging task of EEL.

5.4 Qualitative Analysis

To complement our quantitative evaluation, we employ a Large Language Model (LLM) as a judge (Gu et al., 2024) to provide qualitative insights into the quality of generated descriptions. This approach leverages the LLM's ability to assess linguistic fluency, semantic coherence, and relevance without relying on explicit gold references (Bai et al., 2023). By comparing descriptions side-by-side, the LLM determines which one better represents the target entity in the given context, providing win rates for each comparison. For this experiment, we use Phi-3.5-mini-instruct (Abdin et al., 2024) as the judge model. The model is prompted in a pairwise manner to determine which of the two descriptions better aligns with a given mention, using the context of an entity from the unseen splits of the AIDA validation and test sets. These splits contain only entities that were not seen during training. An example of the prompt used in this experiment is shown in Appendix B.

For this experiment, we conduct three sets of comparisons: (1) T5$_{RAED}$ vs. Wikipedia gold descriptions, (2) T5 without retrieval, vs. Wikipedia gold descriptions, and (3) T5$_{RAED}$ vs. T5 without retrieval. The win rates reported in Table 4 reveal that, as expected, the gold descriptions perform best overall, RAED's descriptions are strongly preferred over descriptions without retrieval with a win rate of 82%. Interestingly, RAED descriptions are even competitive with gold descriptions, achieving a 40% win rate, whereas the non-retrieval definitions win in only 9% of cases.

Finally, in Table 5, we present qualitative examples from our analysis of RAED, highlighting recurring error types made by models without retrieval, that RAED effectively avoids. In the first example, the T5 model hallucinates the middle name and birth date of *Julia Roberts*, whereas RAED generates factually accurate information and even includes examples of films she is known for, which are not mentioned in the gold definition. In the second example, the model trained without retrieved passages incorrectly states that *Benghazi* is the capital and largest city of Libya. In contrast, RAED, using the same underlying generator, correctly identifies it as the second-largest city and also names the actual capital, *Tripoli*. The third example illustrates a disambiguation error: T5 fails to resolve the mention correctly and mistakenly generates a definition for *Japan* instead of the intended entity, the *Japan national football team*. T5$_{RAED}$, on the other hand, correctly disambiguates the entity in question and generates an accurate definition.

6 Conclusions

In this work, we introduce RAED, a retrieval-augmented framework for generating entity ti-

Error	**Factual**
Input	*[...] The film starring Liam Neeson and <u>Julia Roberts</u> recounts the life of Michael [...]*
T5	Julia Roberts <def> Julia `Anne` Roberts (`born November 17, 1967`) is an American actress and filmmaker.
T5$_{RAED}$	Julia Roberts <def> Julia `Fiona` Roberts (`born October 28, 1967`) is an American actress `known for her roles in films such as "Pretty Woman" and "Erin Brockovich".`
Gold	Julia Roberts <def> Julia Fiona Roberts (born October 28, 1967) is an American actress.

Error	**Factual**
Input	*[...] police officers clash regularly in <u>Benghazi</u> At least [...]*
T5	Benghazi <def> `Benghazi is the capital and largest city of Libya.`
T5$_{RAED}$	Benghazi <def> `Benghazi is the second-largest city in Libya, after the capital Tripoli.`
Gold	Benghazi <def> Benghazi is the second-most-populous city in Libya as well as the largest city in Cyrenaica, with an estimated population of 859,000 in 2023.

Error	**Disambiguation**
Input	*Soccer Late Goals Give <u>Japan</u> Win Over Syria [...]*
T5	`Japan` <def> Japan is an island country in East Asia.
T5$_{RAED}$	`Japan national football team` <def> The Japan national football team represents Japan in men's international football and is controlled by the Japan Football Association, the governing body for football in Japan.
Gold	Japan national football team <def> The Japan national football team, also known by the nickname Samurai Blue, represents Japan in men's international football.

Error	**Disambiguation**
Input	*[...] although sixth seeded former champion <u>Agassi</u> had to wriggle out of a dangerous [...]*
T5	`Michael Agassi` <def> `Michael David Agassi (born June 29, 1970)` is an American former world No. 1 tennis player.
T5$_{RAED}$	`Andre Agassi` <def> `Andre Kirk Agassi (born April 29, 1970)` is an American former world No. 1 tennis player.
Gold	Andre Agassi <def> Andre Kirk Agassi (born April 29, 1970) is an American former world No. 1 tennis player.

Table 5: Qualitative examples comparing descriptions generated by RAED and the model without retrieval. For each example, we indicate the type of error, highlighting incorrect parts in red and corrected sections in green.

tles and definitions, i.e., descriptions, to address the challenge of emerging entities in knowledge-intensive NLP tasks. Unlike traditional disambiguation systems that rely on predefined knowledge bases, RAED retrieves relevant passages to generate informative descriptions, enabling generation even for entities missing from inventories.

Through extensive evaluations, we show that RAED effectively combines retrieval and generation to produce high-quality descriptions, significantly improving entity disambiguation performance. Our results on TempEL highlight RAED's ability to handle emerging entities, bridging the gap between static knowledge bases and evolving entity representations. Moreover, our qualitative

evaluation using an LLM-as-a-Judge further supports the advantages of RAED over context-only approaches.

Despite these advancements, challenges remain. The gap between generated and human-curated definitions suggests that future work should explore scaling RAED using larger datasets to enhance adaptability to emerging knowledge. Another key open problem is identifying emerging entities, i.e., entity discovery, rather than only describing them post-hoc. To address this, future efforts could focus on integrating entity discovery, retrieval, and generation into a unified framework. This would improve robustness in knowledge base completion and knowledge graph construction.

7 Limitations

RAED relies solely on Wikipedia as a knowledge source, which, despite its coverage, is incomplete and may contain outdated or biased information. This limits RAED's ability to describe entities that are poorly covered or emerging outside of mainstream documentation. Incorporating additional sources such as news archives or domain-specific corpora could improve robustness.

Another limitation is the system's reliance on retrieval quality. If the retriever fails to find relevant passages, the generated descriptions may be inaccurate or incomplete. While retrieval reduces hallucinations compared to purely generative models, errors in retrieved content can propagate through the generation process.

Finally, RAED does not proactively identify new entities but rather generates descriptions for given mentions. Future work could explore integrating entity discovery into the RAED to enable dynamic detection of emerging entities.

Acknowledgements

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

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

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems*, volume 36, pages 78142–78167. Curran Associates, Inc.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

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.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Nora Kassner, Fabio Petroni, Mikhail Plekhanov, Sebastian Riedel, and Nicola Cancedda. 2022. EDIN: An end-to-end benchmark and pipeline for unknown entity discovery and indexing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8659–8673, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ben Kurtovic. 2013. parserfromhell. https://github.com/earwig/mwparserfromhell.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Luigi Procopio, Simone Conia, Edoardo Barba, and Roberto Navigli. 2023. Entity disambiguation with entity definitions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1297–1303, Dubrovnik, Croatia. Association for Computational Linguistics.

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(140):1–67.

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.

Alessandro Scirè, Andrei Stefan Bejgu, Simone Tedeschi, Karim Ghonim, Federico Martelli, and Roberto Navigli. 2024a. Truth or mirage? towards end-to-end factuality evaluation with llm-oasis. *arXiv preprint arXiv:2411.19655*. To appear in Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024b. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Junxiong Wang, Ali Mousavi, Omar Attia, Ronak Pradeep, Saloni Potdar, Alexander Rush, Umar Farooq Minhas, and Yunyao Li. 2024a. Entity disambiguation via fusion entity decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6524–6536, Mexico City, Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

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.

Klim Zaporojets, Lucie-Aimée Kaffee, Johannes Deleu, Thomas Demeester, Chris Develder, and Isabelle Augenstein. 2022. Tempel: Linking dynamically evolving and newly emerging entities. In *Advances in*

Model	Desc	2020	2021	2022
	Gold	71.0	64.8	72.3
2019	T5	51.2	37.9	44.5
	T5$_{RAED}$	**57.0**	**47.4**	**53.5**

Table 6: We evaluate EL performance on temporal test sets (2020–2022) using descriptions generated by T5-Large, RAED (with T5-Large), and the gold description. Results in bold indicate the best performance for generated descriptions.

Neural Information Processing Systems, volume 35, pages 1850–1866. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

A Entity Description Generation Details

For decoder-only models, an example of a prompt used to generate the entity description (title and definition) is shown in Table 7.

B LLM-as-a-Judge Details

An example of a prompt used for the LLM-as-a-Judge experiment is shown in Table 8. We use Phi-3.5-mini-instruct[11] (Abdin et al., 2024) as the judge model.

C Impact of number of retrieved passages

Table 9 presents the InKB Micro F1 scores of our T5$_{RAED}$ model using varying numbers of retrieved passages. The number of passages indicated for each row is used during both training and evaluation. Table 9 shows that T5 consistently benefits from additional retrieved context, even when using only three passages at training and inference time. We observe steady performance improvements up to ten passages. Beyond this point, however, performance begins to degrade, particularly at 20 passages. We hypothesize that this decline is due to the increasing noise introduced by excessive context, which becomes difficult for the model to handle in this challenging disambiguation task.

---

[11]https://huggingface.co/microsoft/Phi-3.5-mini-instruct

Generate a title and definition seperated by <def> for the mention of an entity between [DEFINE] and [/DEFINE].
Input Text:
INPUT_TEXT
Passages:
Passage 1 ...
Passage 2 ...
Passage 3 ...
Passage 4 ...
Passage 5 ...
Target Entity Description:
"COVID-19 is a contagious disease caused by the coronavirus SARS-CoV-2".

Table 7: Prompt for Entity Description Generation by RAED.

Given the context:
INPUT_TEXT
which entity definition better matches the highlighted mention for entity *"WIKIPEDIA_PAGE_TITLE"*?
Description 1: 'Desc_1'
Description 2: 'Desc_2'
Answer 'Description 1' or 'Description 2'."

Table 8: Prompt used for our LLM-as-Judge experiments. It is important to note that the order in which the two descriptions appear is random.

D Emerging Entity Linking Details

In this section, we provide additional details and analysis of the Emerging Entity Linking experiment. TempEL (Zaporojets et al., 2022) offers not only temporal splits of Wikipedia but also Entity Linking models trained on each yearly snapshot. For instance, TempEL provides ten bi-encoder models trained on Wikipedia snapshots from 2013 to 2022, meaning that the 2019 TempEL model was trained solely on the 2019 Wikipedia split. We use this model in our experiments to accurately simulate the emerging entity setting and to avoid data leakage from model training. Table 6 shows not only the performance of the TempEL model when provided with the generated definitions in its index, but also with the gold definitions. This would be considered the ceiling for performance in Entity Description for Emerging Entity Linking.

Model	Passages	In-domain	Out-of-domain					AVG	
		AIDA-B	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	ALL	OOD
T5	✗	90.1	91.3	78.4	89.4	74.4	72.2	82.6	81.1
T5$_{RAED}$	3	91.3	91.3	80.2	88.3	74.2	74.4	83.3	81.7
T5$_{RAED}$	5	91.6	90.4	81.6	88.3	74.2	73.8	83.3	81.6
T5$_{RAED}$	10	**91.7**	**92.3**	82.0	90.7	**77.3**	**78.7**	**85.5**	**84.2**
T5$_{RAED}$	20	89.4	91.3	**84.0**	**91.4**	76.0	75.1	84.5	83.5

Table 9: InKB Micro F1 comparison of our T5$_{RAED}$ model with varying numbers of retrieved passages. The 0-passage line represents the baseline, while the 10-passage line corresponds to the results in Table 2. Results in bold indicate the best performance.