Integrating Large Language Models in Multimodal Entity Linking: A Novel Two-Level Reflection Framework

Anonymous ACL submission

Abstract

Multimodal Entity Linking (MEL) is an essential technology in numerous applications. Existing methods depend on designing complex multimodal interaction modules and require extensive domain-specific training data. As the traditional pretrain-finetune paradigm evolves towards prompt engineering with large language models (LLMs), investigating prompt engineering-based MEL approaches becomes increasingly vital. However, using LLMs with straightforward instructions presents challenges in MEL tasks. These include context-unfaithful fine-grained entity selection and the overlooking of key details due to information overload. To this end, this paper introduces a novel two-level reflection framework for MEL tasks, named SMCR. In this framework, an LLM is used for entity selection. To address context-unfaithfulness, we implement semantic consistency reflection based on LLM's self-feedback. To simplify the complexity of image utilization and alleviate information overload, we introduce modality consistency reflection. This approach iteratively integrates visual clues through external feedback. Experimental results on two established public MEL datasets show that our solution achieves state-of-the-art performance. Further analysis confirms the effectiveness of our proposed modules. Our code is available at `https://anonymous.4open.science/r/SMCR-1215`.

1 Introduction

Entity linking, the task of mapping ambiguous mentions in text to standard entities in a given knowledge base (KB, e.g., Wikipedia) (Shen et al., 2014). It serves as a pivotal technology in various applications including knowledge graph population (Lin et al., 2020), question answering (Shah et al., 2019; Longpre et al., 2021), and recommendation systems (Deldjoo et al., 2020). Given the prevalence of multimodal contexts (images and texts) in real-world scenarios, recent studies (Wang et al., 2022b; Yao et al., 2023) suggest incorporating images to enhance entity disambiguation, leading to the emergence of Multimodal Entity Linking (MEL).

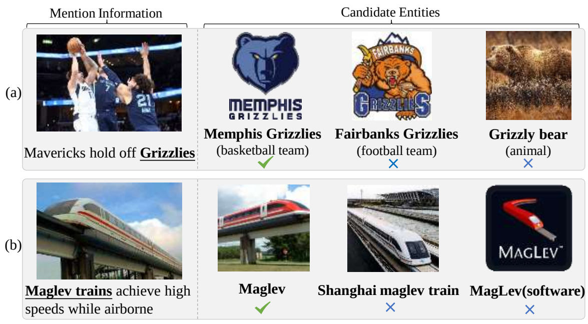

Figure 1: Typical examples of the MEL Task. (a) Images play a crucial role in disambiguation; (b) A bad case demonstrating fine-grained hallucinations in large language models.

Existing MEL methods are all based on the pretrain-finetune paradigm, often requiring complex multimodal interaction modules for feature extraction (Dongjie and Huang, 2022; Luo et al., 2023) or additional domain-specific pretraining data (Wang et al., 2023). This poses significant barriers in practical applications. With the emergence of Large Language Models (LLMs, e.g., ChatGPT), an increasing body of research (Zhao et al., 2023; Chen et al., 2023) demonstrates their exceptional performance in knowledge-intensive tasks. Thus, employing LLMs with several demonstrations as alternatives to traditional methods has emerged as a practical solution for various tasks. Exploring prompt engineering-based MEL methods holds critical importance.

However, employing existing LLMs for MEL tasks presents several challenges. Firstly, these models often produce hallucinations that are not contextually grounded. For instance, as illustrated in Figure 1 (b), the mention "Maglev trains" should link to the entity "Maglev". However, between "Maglev" and "Shanghai maglev train", LLMs tend to select the more specific entity "Shanghai maglev train", despite the absence of supporting context. Secondly, there's the issue of information overload. For the mention images, we employ a series of image-to-text models to generate multi-faceted textual descriptions. Feeding all these descriptions to the LLM simultaneously imposes a significant information burden (Xi et al., 2023), causing the LLMs to overlook critical information and make incorrect inferences.

To address the above problems, this paper introduces an innovative approach known as LLM-based **S**emantic and **M**odality **C**onsistency **R**eflections (denoted as SMCR) for MEL task. Initially, we adopt an LLM (e.g., GPT-3.5), with carefully crafted prompts to select a candidate entity from the KB for a given mention. Subsequently, semantic consistency reflection is designed to evaluate the semantic granularity between the entity and mention, thereby determining the necessity of re-selection. Finally, the approach introduces a modality consistency reflection, involving inter-modal consistency verification and visual iterative feedback, to decide if further selection based on visual clues is required. Our method effectively addresses the aforementioned challenges through three key characteristics. 1) Semantic Consistency Reflection. Direct selection without verification may lead to results unfaithful to the context. We emphasize the LLM's focus on mention context for choosing entities through semantic consistency reflection. 2) Inter-Modal Consistency Verification. We propose an innovative utilization of images. Initially, the LLM selects candidate entities based on textual modality, then uses the visual modality for verification. This approach, as opposed to combining text and image modalities for selection, simplifies the task and reduces the noise inputted to the LLM, allowing it to concentrate solely on the textual context, while leaving complex image information to specialized models (e.g., CLIP (Radford et al., 2021)). 3) Visual Iterative Feedback. In scenarios necessitating image clues, we employ four rounds of iteration invoking various image-to-text models, fully exploiting images from diverse perspectives and avoiding information overload.

Contributions. The contributions of this paper are summarized as follows:

- We propose a novel approach for image utilization. Using visual modality to verify textual results and iteratively integrating image clues when text clues are partially absent. This method simplifies the complexity of fusing image and text information.

- We present the SMCR framework, designed to address the issues of context-unfaithfulness and information overload encountered in LLMs when applied to MEL tasks. To our knowledge, this is the first work to propose prompting LLMs for MEL tasks.

- Experimental results show that our model achieves state-of-the-art performance, attaining a top-1 accuracy of 90.58% (+ 2.6%) on WikiMEL and 80.57% (+ 1.5%) on WikiDiverse. Notably, our method requires no training and is easily transferable.

2 Related Works

Multimodal Entity Linking. The existing works can be divided into two categories: 1) Similarity-ranking based entity linking (Gan et al., 2021; Wang et al., 2022a; Yao et al., 2023) and 2) Generative entity linking (De Cao et al., 2020; Wang et al., 2023). The first category involves a two-step process. Initially conducting candidate retrieval (Yamada et al., 2016; Ganea and Hofmann, 2017) to obtain a set of top-k candidate entities closest to the mention, followed by entity re-ranking. These methods focus on learning the multimodal features of mentions and entities. For instance, Wang et al., 2022a employ co-attention at both token and phrase levels to construct visual-guided textual features and textual-guided visual features, ultimately obtaining a joint multimodal representation through gated fusion. Typically, the similarity between entities and mentions is simply obtained through the cosine similarity (Wang et al., 2022b). Considering the topical coherence of mentions appearing in the same context, some studies (Le and Titov, 2018; Yang et al., 2023a) propose joint disambiguation for multiple mentions. However, this type of method requires designing complex multimodal interaction modules. Meanwhile, the context of a mention may not precisely describe the mention itself, posing challenges in learning its multimodal features. The second category centers on training generative language models to encode the multimodal context of mentions. Target entity names are directly decoded using constrained generation (De Cao et al., 2020) techniques. This demands profound background knowledge, necessitating extensive domain-specific training data. For example, Wang et al., 2023, collected additional multimodal data from BLINK and Wikipedia KB for pretraining.

LLM-based Reflection. Large Language Models (LLMs) have been extensively employed in various NLP tasks. However, their performance is hindered by issues such as hallucinations and unfaithful reasoning. A proposed solution to these challenges involves incorporating reflection steps (Pan et al., 2023). The sources of feedback for reflection are categorized into two types: 1) Self-provided feedback by the LLM (Shinn et al., 2023) and 2) Feedback injected through external means (Peng et al., 2023). The first category leverages the LLM itself for both evaluation and refinement, such as SELFCHECK (Miao et al., 2023) and SELF-REFINE (Madaan et al., 2023). It is typically iterative, continuing until the output meets certain criteria or is interrupted in cases of model stagnation. The second category utilizes various external tools to assess and provide feedback on LLM-generated content, such as separately trained models (Akyürek et al., 2023), additional domain-specific knowledge (Peng et al., 2023), and other tools (Welleck et al., 2022). Feedback through external means offers greater flexibility, introducing information not inherent in LLMs and identifying errors that the LLMs themselves may not detect. In our framework, semantic consistency reflection falls under the first category. Modality consistency reflection, where external feedback mechanisms infuse visual information into LLMs, is an example of the second category.

3 Overview

In this section, we first formalize the task of multimodal entity linking and then outline our proposed framework for the task.

3.1 Task Formulation

Multimodal entity linking is the task of aligning mentions within multimodal contexts to their respective entities in a KB. Formally, given $\{(m, T_m, I_m)\}$, where m denotes a mention, T_m is the textual context surrounding m, and I_m is the image context for m, MEL aims to predict a standard entity for each mention: (m, e) $(e \in \mathcal{E})$, where $\mathcal{E}$ is the entity set in the KB.

3.2 Framework

As depicted in Figure 2, our framework mainly has the following four steps: 1) Target Entity Selection. 2) Semantic Consistency Reflection (SCR). 3) Inter-Modal Consistency Verification. 4) Visual Iterative Feedback. Steps 3) and 4) together form the Modality Consistency Reflection (MCR).

- **Target Entity Selection.** With refined one-shot CoT, we employ a large language model (e.g., GPT-3.5-turbo) to select the most probable candidate entity from KB for the mention.
- **Semantic Consistency Reflection.** For the entity selected in step 1, we continue utilizing the LLM, in conjunction with constrastive CoT, to verify its semantic consistency with the mention in its original context, and determining whether a reselection of the candidate entity is warranted.
- **Inter-Modal Consistency Verification.** For the selected entity that passes step 2, we further check its consistency with the mention image. If consistent, it is outputted as the final result.
- **Visual Iterative Feedback.** If the selected entity does not align with the mention image, we extract information from the image and feed it back to step 1. Then, combining this visual feedback, we reselect the candidate entity and initiate a new iteration cycle. In each iteration, we gradually leverage different facets of the image information to prevent information overload.

4 Methodology

In this section, we provide the details of the four key steps involved in our SMCR framework.

4.1 Target Entity Selection

Given a mention and mention context, the purpose of this step is to select a candidate entity for the mention from the KB. In this paper, we employ an LLM (e.g., GPT-3.5-turbo) with ICL to achieve this purpose. Specifically, we first follow existing work (Wang et al., 2022b,a) to retrieve Top-K candidate entities along with their descriptions from the KB (e.g., Wikipedia). Based on the mention, mention context, and candidate entities with descriptions, we construct the input for LLM. This input comprises four components: instructions, ICL, the data

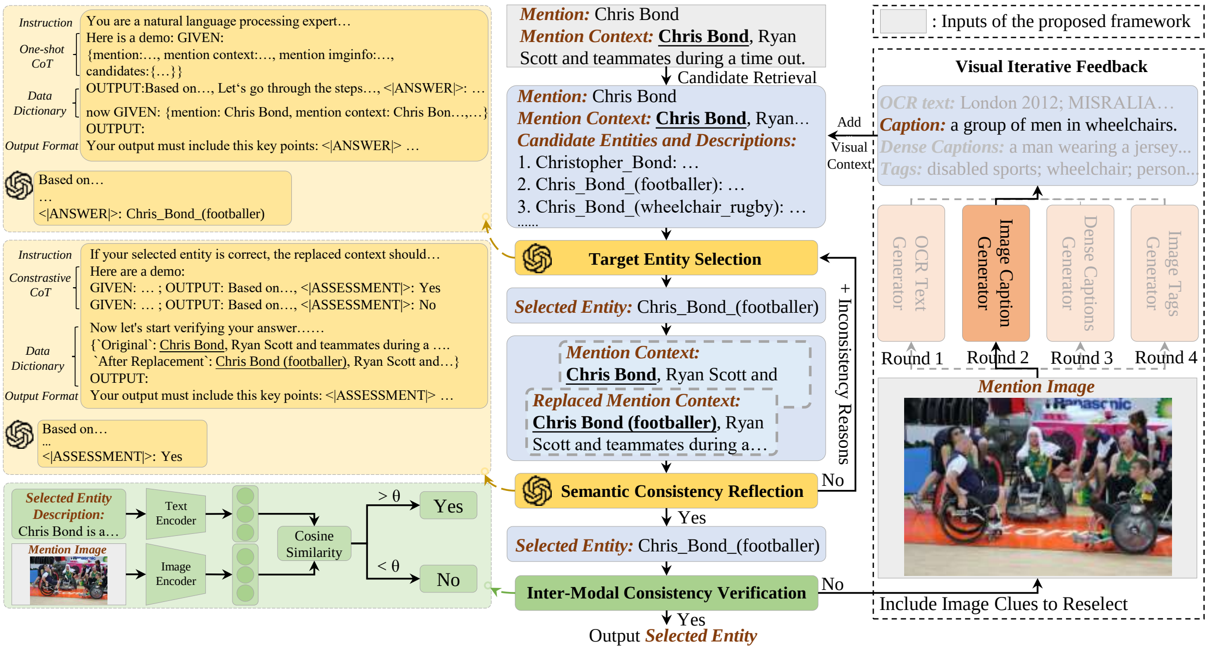

Figure 2: Our framework consists of four key steps. (1) Target Entity Selection. (2) Semantic Consistency Reflection (SCR). (3) Inter-Modal Consistency Verification. and (4) Visual Iterative Feedback. Steps (3) and (4) together form the Modality Consistency Reflection (MCR). The left column shows the details of each step.

dictionary, and the output format specification. In the instruction, we provide the task role and the definition of the multimodal entity linking task. In the ICL, we employ a one-shot CoT (Wei et al., 2022) as an example to demonstrate the steps of reasoning, where the CoT is initially generated by the LLM and then manually refined. CoT consists of three steps: 1) Analyze the mention and context, 2) Compare the mention with each candidate entity, and 3) Select the most relevant candidate entities. The sample data is presented in a dictionary format, with keys including "mention", "mention context", and "candidate entities". Finally, we specify the output format, i.e., "⟨|ANSWER|⟩:⟨your answer⟩" for the LLM, where "⟨your answer⟩" is selected from the candidate entities. Upon inputting this input into the LLM, we obtain the candidate entity from its response.

The purely textual input described above is only used in the initial execution of this step. In subsequent iterations, image information is integrated to assist the LLM in selecting candidate entities. The integration of text and image inputs differs from text-only inputs in two aspects: Firstly, in the CoT, an additional step utilizing visual information is inserted following the first step, titled "Analyze the mention image information and identify helpful details". Secondly, in the data dictionary, we add a new key, i.e., "mention imginfo".

4.2 Semantic Consistency Reflection

This step aims to determine whether the candidate entity identified in the previous step aligns with the mention at the textual semantic level. If there is consistency, we proceed to the next step; otherwise, we return to the first step to re-select a candidate entity. In this step, we maintain the semantic consistency reflection within the same LLM dialogue window used in the previous step. The continuity of the dialogue window provides contextual information beneficial for this task, enhancing model performance.

More specifically, we first replace the mention in its original context with the selected entity to obtain a *Replaced Mention Context*. Then, given both *Mention Context* and *Replaced Mention Context*, we construct the input for LLM, maintaining the same components as in 4.1. It is important to specifically note that in the ICL, we provide *constrastive CoT* for both consistency and inconsistency assessments. Finally, we feed this input into the LLM to analyze whether the semantics remain consistent before and after the replacement. When the assessment is "YES" (signifying semantics consistency), we move forward to the next step. If not, the reasons for inconsistency are added to the historical dialogue record, and we repeat the target entity selection process. This iterative approach continues until the selected entity is verified as consistent, or it reaches a predefined loop limit. In the latter scenario, the last selected entity is chosen as the output.

4.3 Inter-Modal Consistency Verification

For the selected entity that passes SCR, we further assess its alignment with the mention image through Inter-Modal Consistency Verification. If it passes the verification, this entity is then output as the final result; otherwise, we proceed to the next step, incorporating image information for further entity selection.

Given the description D_e of the selected candidate entity e and the mention image I_m, we first encoder them into vectors using the text and image encoders of the CLIP model (Radford et al., 2021). Then, we employ a dot product to compute the cosine similarity between the above two vectors. Finally, we establish a predefined threshold to determine whether the selected entity aligns with the mention image. The above process is formulated as:

$$score(D_e, I_m) = Enc_T(D_e) \cdot Enc_I(I_m), \quad (1)$$

$$assessment = \begin{cases} 1 & score(D_e, I_m) > \theta \\ 0 & score(D_e, I_m) < \theta \end{cases} \quad (2)$$

Here, Enc_T and Enc_I represent the text and image encoders, respectively, and θ is the pre-defined threshold. If the assessment is "1"(YES), the selected entity is output as the final result. Otherwise, we proceed to the next step.

4.4 Visual Iterative Feedback

In response to a "NO" output from the previous step, we incorporate visual information to refine our selection of the entity. This paper utilizes various image-to-text models to generate multi-faceted descriptions for a given image, which include OCR text, image captions, dense captions, and image tags. To prevent information overload, we iteratively apply these different types of descriptions.

Specifically, upon inputting mention image, an image-to-text model is initially invoked to generate an image description (e.g., "a group of men in wheelchairs..."). This description is then integrated as additional visual context into step 1, as detailed in Section 4.1. Subsequently, we execute step 1 again to re-select an entity, thereby initiating a new iteration cycle. During this cycle, we continue to use Inter-Modal Consistency Verification to assess if the selected entity aligns with mention image, deciding whether to utilize other facets of image clues. We employ four distinct models — "OCR", "Image Captioning", "Dense Captioning", and "Image Tagging" — in a specific sequence determined on the WikiDiverse validation set, iterating up to four rounds. If the entity still fails the Inter-Modal Consistency Verification after all iterations, we revert to the entity initially selected.

Table 1: The statistics of datasets.

Dataset	Train	Valid	Test
WikiMEL	18,092	2,585	5,169
WikiDiverse	13,205	1,552	1,570

5 Experiments

In this section, we conduct comprehensive experiments to evaluate our proposed method on two widely-recognized public MEL datasets. Furthermore, extensive analyses are presented to offer deeper understanding of the framework.

5.1 Experimental Setup

Datasets. In this study, we employ two datasets, namely WikiMEL(Wang et al., 2022a) and WikiDiverse(Wang et al., 2022b) for evaluation. WikiMEL collects data from Wikipedia's entity pages, with its primary entity type being Person. It uses Wikidata as its target KB. We followed the original provided method (Wang et al., 2022a) for candidate retrieval. Wikidiverse is built by Wikinews, covering 7 types of entities(i.e., Person, Organization, Location, Country, Event, Works, and Misc). It utilizes Wikipedia as its target KB. Following existing work (Wang et al., 2023), we conduct experiments using the top-10 candidate entities provided by the dataset, and assign the label "nil" when the mention's target entity is not included in the candidate set. The statistics of two datasets are concluded in Table 1. We use the same test set as existing works for evaluation.

Baseline. We compare our proposed method with various state-of-the-art (SOTA) methods, which are divided into two groups: (1) Text-only methods, which include BERT (Kenton and Toutanova, 2019), BLINK (Wu et al., 2020), and GPT-3.5-Turbo[1]. (2) Visual-text fusion methods, which include CLIP (Radford et al., 2021), DZMNED

[1] https://platform.openai.com/docs/models/gpt-3-5

Table 2: Main results on WikiMEL and WikiDiverse. The values in "()" indicate the standard deviation of the results.

Model	Top-1 Accuracy (%)	
	WikiMEL	WikiDiverse
Text		
BERT	31.7	69.6
BLINK	30.8	70.9
GPT-3.5-Turbo	77.1	63.9
GPT-3.5-Turbo (CoT)	79.9	77.1
Text+Vision		
CLIP	79.8	50.5
DZMNED	78.8	56.9
GHMFC	43.6	46.0
LXMERT	20.6	78.6
DRIN	65.5	51.1
MMEL	71.5	-
GDMM-base	68.0	79.1
GDMM-large	72.4	78.7
MIMIC	88.0	63.5
SMCR	**90.58** (0.23)	**80.57** (0.69)

Table 3: The Ablation Study of SMCR on WikiMEL and WikiDiverse. 4.2, 4.3, 4.4 correspond to the sections in this paper.

Model	Top-1 Accuracy (%)	
	WikiMEL	WikiDiverse
SMCR	**90.58** (0.23)	**80.57** (0.69)
w/o CoT	88.20	66.18
w/o 4.2	87.48	79.30
w/o 4.3	86.65	78.22
w/o 4.4	86.26	77.96
w/o 4.2, 4.3	81.51	77.32
w/o 4.2, 4.4	79.96	77.07
w/o 4.3, 4.4	86.26	77.96
w/o 4.2, 4.3, 4.4	79.94	77.07

(Moon et al., 2018), LXMERT (Wang et al., 2022b), GHMFC (Wang et al., 2022a), GDMM(base/large) (Wang et al., 2023), MMEL (Yang et al., 2023a), MIMIC (Luo et al., 2023) and DRIN (Xing et al., 2023).

Metrics. Following existing works (Wang et al., 2022a; Yang et al., 2023a), our evaluation employs the Top-1 accuracy metric.

Implementations. Within the applied framework, we utilize the OpenAI API, specifying the model as "gpt-3.5-turbo-16k-0613", with the temperature set to 0 and other parameters remaining at their default settings. We employ the same One-shot CoT and Contrastive CoT across all samples. To ensure reliability in our results, we conduct three repeated experiments and calculate the standard deviation. For the candidate retrieval in section 4.1, we set $k = 10$. The CLIP model used in Section 4.3 is referred to as CLIP_ViT_bigG_14_laion2B_39B_b160k. We set the θ to 29 based on the WikiDiverse validation set and apply it across all datasets. In Section 4.4, for the applied image-to-text models, we reference existing work (Yang et al., 2023b), employing the latest models from Azure Cognitive Services APIs[2], including Image Captioning, Dense Captioning, Image Tagging, and OCR models.

[2]https://portal.azure.com/#view/Microsoft_Azure_Project-Oxford/CognitiveServicesHub/~/ComputerVision

5.2 Main Results

In this section, we present a comparative analysis of our proposed method against all baseline approaches on WIKIMEL and WikiDiverse datasets. The results are detailed in Table 2.

Based on the experimental results, we can draw the following observations and conclusions. 1) Without any component training, our method outperforms the current state-of-the-art (SOTA) approaches on two datasets, demonstrating the effectiveness of our method. Specifically, on WikiMEL and WikiDiverse, we achieve the top-1 accuracy of 90.58% and 80.57%, respectively, marking improvements of 2.6% and 1.5% over previous SOTA methods. 2) The proposed framework significantly enhances LLM performance in the MEL task, particularly evident in SMCR's significant improvements (13.5% and 16.7%) over the direct application of GPT-3.5-Turbo. 3) Compared to the WikiDiverse (80.57%), our method performs better on WikiMEL (90.58%). This is due to the greater prevalence of "nil" target labels in WikiDiverse, making it a more challenging task to infer the "nil" than identifying the correct entity. 4) The "GPT-3.5-Turbo + CoT" method, using only textual modality, already achieves high accuracy scores on both datasets. This reaffirms our perspective that in the MEL tasks, information provided by the textual modality is predominant. Mention images typically strengthen textual information, yet they serve to supplement missing clues in rare instances.

5.3 Ablation Experiment

This section presents comprehensive ablation studies to validate the effectiveness of each component in our proposed framework. Firstly, we performed ablations on the key steps of the framework, with

Table 4: The Ablation Study on the image-to-text models presented in Section 4.4. (ocr: OCR text, cap: Caption, den: Dense Captions, tag: Tags)

Model	**Top-1 Accuracy (%)**	
	WikiMEL	WikiDiverse
SMCR	**90.58** (0.23)	**80.57** (0.69)
w/o ocr	90.25	79.94
w/o cap	90.23	80.38
w/o den	90.52	80.45
w/o tag	90.38	80.51
w/o ocr, cap	89.77	79.75
w/o ocr, den	90.08	80.06
w/o ocr, tag	89.92	80.19
w/o cap, den	89.77	80.38
w/o cap, tag	89.84	80.38
w/o den, tag	90.25	80.45
w/o ocr, cap, den	88.90	79.87
w/o ocr, cap, tag	89.07	79.75
w/o ocr, den, tag	89.50	80.06
w/o cap, den, tag	88.93	79.87
w/o all	86.26	77.96

the results presented in Table 3. These results show that removing any step led to a decline in model performance, thereby demonstrating the effectiveness of all steps in our framework. Subsequently, ablations were conducted on the four image-to-text models in Step 4.4, summarized in Table 4. All four models utilized in this step contributed positively to the iterative process.

5.4 Detailed Analysis

In this section, we analyze the important components within our framework in detail with in-depth case study.

Improvements analysis for SCR. To investigate the error types effectively mitigated by SCR, we analyzed improved samples from the WikiDiverse test set after SCR integration, as shown in Figure 3, categorizing them into four error types: 1) Fine-Grained Hallucination. In the absence of supporting contextual information, the LLM selects an erroneous entity with finer granularity. 2) Blurred Span. The LLM fails to focus distinctly on the mention's span, resulting in either span expansion or misplaced attention. 3) Part of Speech Confusion. The selected entity misaligns with the mention's grammatical role in the text. 4) Others. Other scenarios of noted improvement. We provide cases for the first three types of errors in Figure 5.

What visual clues does our framework show effective improvement ? We analyzed 200 random samples from the WikiDiverse test set. Following

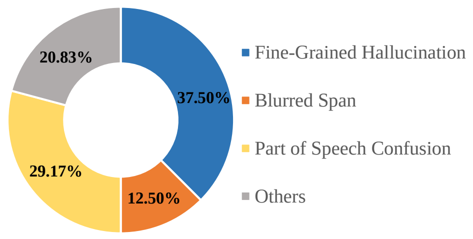

Figure 3: Improvements Decomposition for SCR.

Wang et al., 2022b; Li et al., 2023, we categorize the visual clues into three types: 1) Object: images showing the entity directly, 2) Scene & Property: images depicting associated environments or properties, and 3) Others: additional significant clues. Examples of the first two types are in Figure 4. As shown in Table 5, we observe: 1) Compared to the one-time infusion of all image information (w/o VIF), the iterative use of images shows a primary improvement in Scene & Property. This might be due to the iterative method highlighting finer-grained clues. 2) In comparison to scenarios without visual (w/o Visual), SMCR perform better on Object clues. This underscores our method's efficacy in employing images.

Visual clues	Object	Scene & Property
Image		
Mention Context	A **Shadow** is prepared for flight over Iraq.	Bathum coming to a stop following his **downhill** ride.
Pred (T)	Shadow	Downhill mountain biking
Pred (T+V) = GT	AAI RQ-7 Shadow	Downhill (ski competition)

Figure 4: Examples of the two types of visual clues.

Table 5: Model performance under different visual clues. (w/o VIF: utilizing images Without Visual Iterative Feedback, w/o Visual: Without using images, a: Object, b: Scene Property, c: Others)

Model	**Top-1 Accuracy (%)**			
	a (54)	b (109)	c (37)	total (200)
SMCR	**87.04**	**82.57**	75.68	**82.50**
w/o VIF	85.19	76.15	**78.38**	79.00
w/o Visual	79.63	78.90	75.68	78.50

Efficacy of visual iterative feedback in mitigating information overload. To thoroughly investigate the effects of iterative use of images, we conduct experiments on the WikiDiverse validation set. The results are shown in Figure 6. "Round 0-4"

Error Type	Fine-Grained Hallucination		Blurred Span		Part of Speech Confusion
Image					
Mention Context	**Maglev trains** can accelerate to high speeds as they run suspended in the air	Pujols hit a home run in Sunday's baseball game between the **Anaheim Angels** and the Toronto Blue Jays.	**Egyptian** army soldiers monitor protests over the weekend.	Bart writing "HDTV is worth every cent" in the "**chalkboard gag**.".	An **Iraqi** competitor and an unnamed member of the United States delegation chat.
Pred (w/o SCR)	Shanghai maglev train	1997 Anaheim Angels season	Egyptian Army	Bart Simpson	Iraqis
Pred (w SCR) **= GT**	Maglev	Los Angeles Angels	Egypt	The Simpsons opening sequence	Iraq

Figure 5: Three types of error cases that can be effectively addressed through semantic consistency reflection.

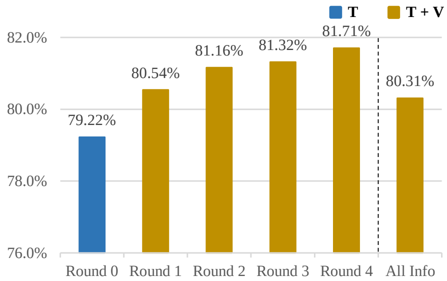

Figure 6: Comparing iterative versus single-use image information processing.

denote the iterative process in our framework and the "All Info" denote a single infusion of images. We calculate the overall Top-1 accuracy after each iteration. From the results, we can see that a one-time infusion of images offers a minimal increase (1.09%), whereas iterative methods yield consistent incremental improvements, demonstrating the efficacy of iterative feedback.

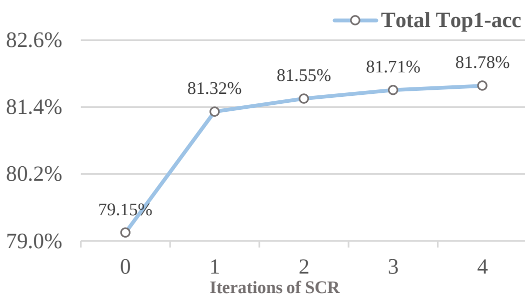

Figure 7: Analysis of convergence iterations for SCR.

Analysis of convergence iterations for SCR. Figure 7 illustrates the convergence iterations of semantic consistency reflection on the WikiDiverse validation set. From the results, two observations can be made: 1) The overall top1-accuracy tends to converge by the third iteration. Therefore, we set the iteration limit of SCR to 3 rounds. 2) The most significant improvement is observed in the first round. This indicates that under the guidance of our framework, the LLM begins to pay significant attention to the mention context for entity selection after making an initial error.

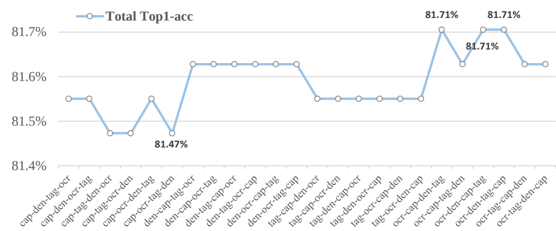

Figure 8: Analyzing the ranking of the four Image-to-Text Models in MCR.

Analyzing the ranking of the four Image-to-Text Models in MCR. Figure 8 illustrates the performance of all permutations of the four image-to-text models applied in Section 4.4 on the WikiDiverse validation set. From the results, we observe that the impact of different permutations on the final results is minimal. Consequently, we simply select the "ocr-cap-den-tag" sequence for implementation.

6 Conclusion

This paper proposes a novel LLM-based two-level reflection framework for the task of MEL. The framework enhances the context-awareness of LLMs through semantic consistency reflection, thereby preventing issues of context-unfaithfulness. The modality consistency reflection specifically facilitates the integration of image and iteratively employs images to alleviate information overload. Experimental results on WikiMEL and WikiDiverse demonstrate that our approach achieves SOTA performance, with additional detailed analyses that validate the effectiveness of each component.

Limitations

The approach of utilizing prompt engineering for multimodal entity linking can be conveniently adapted to practical application scenarios. Despite its advantages, several non-negligible deficiencies persist. Firstly, the utilization of the OpenAI API may encounter limitations in certain scenarios, such as the absence of internet connectivity or constraints imposed by the pricing structure of the API. Additionally, the invocation of the API might raise concerns regarding data confidentiality. Secondly, in real-world scenarios, it's more common for a mention to be absent from the designated Knowledge Base (KB). For such instances of predicting non-existence, there is substantial room for improvement in our method. Lastly, integrating candidate retrieval dynamically with our approach still requires significant effort. We believe that with continued expansion of our framework, it will evolve into a more comprehensive solution in the future.

References

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*.

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

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Zhang Dongjie and Longtao Huang. 2022. Multimodal knowledge learning for named entity disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3160–3169.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: a new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 993–1001.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. *arXiv preprint arXiv:1804.10637*.

Yangning Li, Tingwei Lu, Yinghui Li, Tianyu Yu, Shulin Huang, Hai-Tao Zheng, Rui Zhang, and Jun Yuan. 2023. Mesed: A multi-modal entity set expansion dataset with fine-grained semantic classes and hard negative entities. *arXiv preprint arXiv:2307.14878*.

Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. Kbpearl: a knowledge base population system supported by joint entity and relation linking. *Proceedings of the VLDB Endowment*, 13(7):1035–1049.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1583–1594.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022a. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 938–948.

Sijia Wang, Alexander Hanbo Li, Henry Zhu, Sheng Zhang, Chung-Wei Hang, Pramuditha Perera, Jie Ma, William Wang, Zhiguo Wang, Vittorio Castelli, et al. 2023. Benchmarking diverse-modal entity linking with generative models. *arXiv preprint arXiv:2305.17337*.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022b. Wikidiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. *arXiv preprint arXiv:2204.06347*.

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

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

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.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3599–3608.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. *arXiv preprint arXiv:1601.01343*.

Chengmei Yang, Bowei He, Yimeng Wu, Chao Xing, Lianghua He, and Chen Ma. 2023a. Mmel: a joint learning framework for multi-mention entity linking. In *Uncertainty in Artificial Intelligence*, pages 2411–2421. PMLR.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Barry Menglong Yao, Yu Chen, Qifan Wang, Sijia Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. 2023. Ameli: Enhancing multimodal entity linking with fine-grained attributes. *arXiv preprint arXiv:2305.14725*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.