Enriching Conceptual Knowledge in Language Models through Metaphorical Reference Explanation

Zixuan Zhang and **Heng Ji**
University of Illinois Urbana-Champaign
{zixuan11, hengji}@illinois.edu

Abstract

In this paper, we explore a novel approach to concept enrichment in language models (LMs) by leveraging the fundamental similarities between conceptual knowledge enrichment and metaphorical reference resolution. While previous knowledge editing (KE) methods predominantly focus on factual updates, we introduce a method that trains LMs to not only incorporate new conceptual meanings but also generatively explain the connections between original and enriched definitions through metaphorical analogies. To achieve this, we develop a new dataset tailored for concept enrichment tasks and apply it to train an LM capable of updating and reasoning about conceptual knowledge. The proposed method was evaluated on both "is-a" relation classification and metaphorical reference detection. Experimental results show that our approach significantly enhances the model's ability to understand and apply enriched concepts, demonstrating the potential of metaphorical reference identification in improving conceptual knowledge of LMs[1].

1 Introduction

Large language models (LLMs) demonstrate strong capability in serving as a knowledge system efficient in storing, retrieving, and reasoning across different domains of knowledge (Petroni et al., 2019; Zhao et al., 2022; He et al., 2024). Considering that real-world knowledge is constantly evolving, many research efforts focus on post-training knowledge editing and refinement (Meng et al., 2022, 2023; Liu et al., 2024; Wang et al., 2024; Yu and Ji, 2023; Qin et al., 2024), to ensure that the information in language models remains up-to-date. However, most prior KE research primarily focuses on editing factual knowledge. For example, if the LM knows that *Leonardo DiCaprio* is a citizen of the *United States*, previous KE methods would alter the model to respond with a different country (e.g., *Syria*) when queried about his citizenship. While research in cognitive science (Zhao et al., 2024; Rane et al., 2024) suggests that humans typically grasp new information by learning new concepts, some KE methods also focus on editing concept-level knowledge. Basically when a concept's definition is updated, the edited model should reflect a new understanding of both the concept itself and its related instances.

In this paper, we introduce novel insights by identifying the fundamental similarity between enriching the concepts in LMs and a special case of coreference resolution: *metaphors*. Metaphors, or metaphorical references, typically involve using an existing concept to refer to a new one, where the new and old concepts share significant similarities. For example, the concept of ***stream*** originally referred to a "*body of water with a current flowing within its bed and banks*". However, it now also refers to "*a type of real-time digital transmission of video or audio content*", as both meanings involve the "*continuous flow of some contents*". Almost all metaphor cases are essentially enriching older concepts with new meanings, which closely parallels the task of concept enrichment for LMs.

Based on these similarities, we propose a novel and effective method for enriching conceptual knowledge in LMs by training the model to explain metaphorical references. Specifically, when provided with an updated definition of an old concept, our approach trains the model not only to memorize the new meaning, but also to generatively explain the similarity between the old and new meanings, ensuring that the LM gains a deeper understanding of why the enrichment is valid. We develop a new dataset for the task of LM concept enrichment and use it to train a language model for updating conceptual knowledge. Our model is evaluated on both concept definition memorization

[1]Data and code are available at https://github.com/zhangzx-uiuc/ConceptEnrich.

Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2024), pages 18–22
November 15, 2024.

and sub-instance classification. We also assess its performance on metaphorical reference detection. Experimental results demonstrate the effectiveness of using metaphorical reference generation to enhance LM concept enrichment.

Our contributions can be summarized as follows:

- We propose a new problem setting focused on enriching conceptual knowledge in language models, addressing the realistic need for knowledge to be continuously updated to reflect the dynamic nature of the real world.
- We introduce a novel approach that incorporates metaphorical reference explanation as a training objective, demonstrating its effectiveness both theoretically and empirically.
- We develop and release a new benchmark dataset, *ConceptEnrich*, designed for the task of conceptual knowledge enrichment.

2 Related Work

Conceptual Knowledge Editing Most previous work on knowledge editing in LMs has primarily focused on modifying factual knowledge, with only one prior study, *ConceptEdit* (Wang et al., 2024), addressing the editing of conceptual knowledge in LMs. However, we identify a critical flaw in the basic problem setting of ConceptEdit: the updated concept definitions are often unrealistic, and simply swapped from the definition of other concepts. For example, the model is expected to update the definition of ***stream*** as *a major international multi-sport event* (***Olympics***). We argue that such a setting is not realistic as it never happens in the real world. Additionally, since LMs typically develop understandings of concepts by seeing large amounts of contextual examples during pre-training, an unrealistic edit without providing relevant contexts and examples can break the model's existing knowledge structure, leading to a cascade of related failures in the language model.

Metaphor Detection and Resolution Metaphor detection and resolution have long been central tasks in computational linguistics. With the advent of increasingly powerful language models, researchers have begun to explore how effectively these models can understand metaphors. For instance, (Aghazadeh et al., 2022) investigate the capabilities of current language models in handling metaphors by designing a specific probing task and

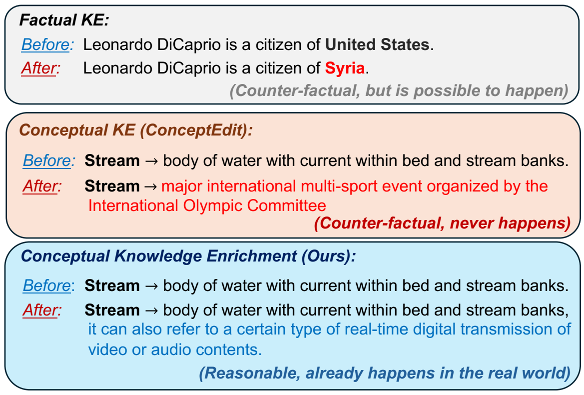

Figure 1: Comparison of the problem settings of traditional factual KE, concept knowledge editing, and concept knowledge enrichment.

dataset. More recently, (Chakrabarty et al., 2023) examined the intersection of visual language models and metaphor detection, evaluating how well diffusion models perform in this complex task.

3 Approach

Problem Formulation We use $p_\theta(\cdot)$ to denote a language model parameterized by θ. Given a set of concepts $\mathcal{C}$, where each concept $c \in \mathcal{C}$ is along with an existing definition $d_{old}(c)$ and a new enriched definition $d_{new}(c)$, our objective is to obtain an updated LM θ_{new} with enriched concept understandings. For example, if c is *tablet*, then $d_{old}(c)$ and $d_{new}(c)$ could be "*a flat piece or slab of stone, clay, wood, or other material, often rectangular in shape, used as a writing surface*" and "*portable touchscreen electronic devices*" respectively.

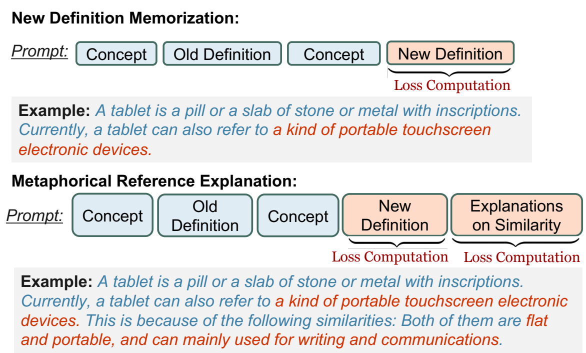

Figure 2: Comparison between the training objectives of *new Definition Memorization* and *Metaphorical Reference Explanation*.

New Definition Memorization We first empower the LM with the fundamental memorization of new definitions of concepts, by directly maximizing the likelihood of new definitions (as illustrated in Figure 2). The loss function can be formulated as a text completion task:

$$\mathcal{L}_{mem}(c) = -\log p_\theta \left(d_{new}(c) \mid c, d_{old}(c)\right). \quad (1)$$

Metaphorical Reference Explanation To further reinforce the model's understanding of the validity of newly enriched concept definitions, we propose a novel method that teaches the model to explain metaphorical references. Specifically, this involves generatively explaining the similarity between the original and new definitions of the concepts. As illustrated in Figure 2, given the concept name and its original definition, we train the model not only to memorize the new definition but also to generate explanations that highlight the similarities between the original and new definitions, clarifying why the enrichment is reasonable. The loss function is formulated as

$$\mathcal{L}_{ref}(c) = -\log p_\theta \left(\textit{sim}(c) \mid c, d_{new}(c), d_{old}(c)\right),$$

where $\textit{sim}(c)$ is a textual description on the similarity between the old definition and the new definition. For example, for the original and enriched definitions of ***tablet***, $\textit{sim}(c)$ could be "*flat and portable, and can mainly used for writing and communications.*" Note that such a similarity description can be obtained from the dataset, or generated by the model itself. We evaluate both of these settings in our experiments. The final training objective is a weighted sum of the two loss values.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{mem} + (1-\alpha) \cdot \mathcal{L}_{ref}.$$

4 Experiments

4.1 Data

ConceptEnrich Previous work (Wang et al., 2024) develops the *ConceptEdit* dataset that contains a series of concepts with their original and edited definitions. However, as discussed in Section 2, we believe that it is not realistic to directly change the definitions of concepts that are completely unrelated. Therefore, in this paper, we develop a new benchmark dataset, *ConceptEnrich*, which contains 121 concepts that are believed to be substantially enriched recently. The dataset is generated with the assistance of GPT-4, where we prompt the model to brainstorm concepts that have acquired enriched meanings in recent years. The model will also generate their old and new definitions, along with a description of their similarities and some typical instances of the concept. The detailed prompt and one generated example are shown in Figure 3.

Metaphor Detection: VUA Corpus Since we conduct conceptual knowledge enrichment for LMs by training the model to generate explanations for metaphorical reference, it would also be interesting to investigate whether the model with enriched concept understandings can be improved in real linguistic metaphor detection tasks. We adopt the widely-used VUA Corpus (Steen et al., 2010) and test whether our model can perform better.

4.2 Evaluation Metrics

For evaluation metrics, similar to (Wang et al., 2024), we mainly focus on whether the a model taught with enriched concept definitions can perform better in classifying its sub-instances. For example, if the model has already known "*stream*" can be extended to "*digital transmission of audio or video content without the need for downloading*", can the model correctly identify "*Twitch*" is a certain kind of stream? For each concept and its sub-instances presented in *ConceptEnrich*, we manually construct the same number of negative examples from the sub-instances from other concepts. Then, we use the model to perform a classification task to identify which instances belong to the concept with an enriched meaning. We compute the AUC of the binary classification task and use it to compare the performances of different models. For metaphor reference detection task, we also compute both the accuracy and AUC of metaphor detection.

4.3 Base Model Setup

In this paper, we adopt GPT2-XL (Radford et al., 2019) as our base LM. We choose to use a model released a few years ago because our primary focus in this paper is to evaluate the model's ability to learn updated definitions of concepts. However, many of the most recent open-source language models already include a wide range of concepts in their pre-training data. To ensure a fair comparison and eliminate the influence of existing prior knowledge, we opted for an older model GPT-2. We also adopt the model with the largest available size to ensure that the base model's capability is still robust and strong enough for our evaluations.

4.4 Main Results

To test the effectiveness of our proposed metaphorical reference explanation approach, we mainly compare our final trained model with the baseline

Can you think of some concepts (in English) whose meanings have been enriched or changed in the last 5 years? For example, previously, the concept "tablet" is defined as a flat piece or slab of stone, clay, wood, or other material, often rectangular in shape, used as a writing surface. But now, "tablet" can also refer to a certain kind of touchscreen electronic devices.

Please generate your answers in the following format:
Concept: Tablet
Old Definition: a flat piece or slab of stone, clay, wood, or other material, often rectangular in shape, used as a writing surface
Enriched Definition: a certain kind of touchscreen electronic devices.
Explanation: 1. Both of these are flat and portable. 2. Both of these enables direct interaction with users. 3. Both of these are mainly used for writing and communications.
Examples: Apple iPad, Microsoft Surface Pro, Amazon Kindle

Concept: Stream

1. **Old Definition**: A small, narrow river.
2. **Enriched Definition**: The digital transmission of audio or video content without the need for downloading.
3. **Explanation**:
   1. Both involve the continuous flow of a medium—water in the physical form, data in the digital form.
   2. Both are accessed and utilized as they move along their path.
   3. Both are natural and intuitive ways to receive resources or information.
4. **Examples**:
   Twitch streaming,
   Netflix video streaming,
   Spotify music streaming

Figure 3: The detailed prompt we use to generate data **(left)** and an example generated example from GPT-4 **(right)**.

Models	Accuracy	AUC
GPT2-XL	55.3	50.0
GPT2-XL + Memorization	61.9	64.8
+ *MetaphorExp* (self-generated)	81.0	85.4
+ *MetaphorExp* (GPT4-generated)	**89.5**	**91.3**

Table 1: Performance (%) for sub-instance classification in our proposed *ConceptEnrich* benchmark.

Models	Accuracy	AUC
GPT2-XL	78.3	82.4
GPT2-XL + Memorization	79.0	83.5
+ *MetaphorExp* (self-generated)	79.3	84.4
+ *MetaphorExp* (GPT4-generated)	**81.9**	**86.3**

Table 2: Performance (%) for metaphorical reference detection on the verb-only subset in the VUA corpus.

model only trained with new definition memorization (*GPT2-XL + Memorization*). Additionally, we evaluate the metaphorical reference explanation approach in both of the following settings: using similarity descriptions from the *ConceptEnrich* dataset (+*MetaphorExp* (GPT4-generated)) and those generated by the model itself (+*MetaphorExp* (self-generated)). This allows us to assess whether our approach is robust enough when no predefined similarity descriptions are provided.

From the results in Table 1, we observe that training the model to memorize only the new definitions of concepts enhances its ability to identify concept sub-instances. Furthermore, our approach, which incorporates metaphorical reference explanations, significantly boosts performance, achieving a 91.3% AUC on the *ConceptEnrich* benchmark. Additionally, even when using self-generated explanations without incorporating any new information, our model still achieves an 85.4% AUC, which is significantly higher than the baseline model that relies solely on memorization. These results demonstrate that using metaphorical reference explanation methods can better help the model to understand and learn enriched meanings of concepts. In Table 2 and Table 3, we can observe similar trends on metaphor detection tasks. These results demonstrate that learning enriched meanings of existing concepts, particularly by exploiting the similarities between old and new definitions, also enhances the language model's ability to detect and understand metaphorical references.

Models	Accuracy	AUC
GPT2-XL	80.3	85.0
GPT2-XL + Memorization	80.5	85.1
+ *MetaphorExp* (self-generated)	81.1	86.0
+ *MetaphorExp* (GPT4-generated)	**85.6**	**89.1**

Table 3: Performance (%) for metaphorical reference detection on the full set of the VUA corpus.

5 Conclusions and Future Work

In this paper, we present a novel and effective approach to concept enrichment in language models by integrating metaphorical reference resolution. The results demonstrate that leveraging metaphorical analogies can significantly enhance a model's ability to comprehend and apply new conceptual knowledge, offering a more nuanced understanding than baseline methods of simply training the model to memorize new concept definitions. The development of a specialized dataset and the successful application of our method to concept sub-instance classification and metaphorical reference detection underscore the potential of our approach.

In future, we plan to explore the scalability of our approach across different domains and languages. Additionally, investigating the integration of our method with other knowledge enrichment techniques, such as continual learning, could further enhance the adaptability and robustness of LMs.

Acknowledgement

We thank the anonymous reviewers for their constructive suggestions. This research is based upon work supported by U.S. DARPA ECOLE Program No. #HR00112390060 and DARPA SemaFor Program No. HR001120C0123. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Qiyuan He, Yizhong Wang, and Wenya Wang. 2024. Can language models act as knowledge bases at scale? *Preprint*, arXiv:2402.14273.

Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, and Heng Ji. 2024. Evedit: Event-based knowledge editing with deductive editing boundaries. *Preprint*, arXiv:2402.11324.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jiaxin Qin, Zixuan Zhang, Chi Han, Manling Li, Pengfei Yu, and Heng Ji. 2024. Why does new knowledge create messy ripple effects in llms? In *arxiv*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sunayana Rane, Polyphony J Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L Griffiths. 2024. Concept alignment. *arXiv preprint arXiv:2401.08672*.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, Tryntje Pasma, et al. 2010. *A method for linguistic metaphor identification*. John Benjamins Publishing Company Amsterdam.

Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Editing conceptual knowledge for large language models. *Preprint*, arXiv:2403.06259.

Pengfei Yu and Heng Ji. 2023. Self information update for large language models through mitigating exposure bias. In *arxiv*.

Bonan Zhao, Christopher G Lucas, and Neil R Bramley. 2024. A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, 8(1):125–136.

Ruilin Zhao, Feng Zhao, Guandong Xu, Sixiao Zhang, and Hai Jin. 2022. Can language models serve as temporal knowledge bases? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2024–2037, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.