

# Editing Conceptual Knowledge for Large Language Models

Xiaohan Wang<sup>♠◇</sup>, Shengyu Mao<sup>♠◇</sup>, Ningyu Zhang<sup>♠◇\*</sup>, Shumin Deng<sup>♠</sup>,  
Yunzhi Yao<sup>♠◇</sup>, Yue Shen<sup>♡</sup>, Lei Liang<sup>♡</sup>, Jinjie Gu<sup>♡</sup>, Huajun Chen<sup>♠◇\*</sup>

<sup>♠</sup>Zhejiang University, <sup>◇</sup>ZJU-Ant Group Joint Research Center for Knowledge Graphs,

<sup>♠</sup>National University of Singapore, <sup>♡</sup>Ant Group

{wangxh07, zhangningyu}@zju.edu.cn

 <https://zjunlp.github.io/project/ConceptEdit>

## Abstract

Recently, there has been a growing interest in knowledge editing for Large Language Models (LLMs). Current approaches and evaluations merely explore instance-level editing, while whether LLMs possess the capability to modify concepts remains unclear. This paper pioneers the investigation of editing conceptual knowledge for LLMs, by constructing a novel benchmark dataset **ConceptEdit** and establishing a suite of new metrics for evaluation. The experimental results reveal that, although existing editing methods can efficiently modify concept-level definitions to some extent, they also have the potential to distort the related instantial knowledge in LLMs, leading to poor performance. We anticipate this work can inspire further progress in understanding LLMs.

## 1 Introduction

The emergence of Large Language Models (LLMs) represents a significant step towards the era of AGI, with the performance of large-scale models being evident for all to see (Bubeck et al., 2023; Zhao et al., 2023c). Despite their advancements, LLMs encounter challenges such as misinformation, outdated knowledge due to the training cut-off, and the risk of producing toxic content (Augenstein et al., 2023; Wang et al., 2023a,e; Zhang et al., 2023; Sun et al., 2024; Feng et al., 2023; Ji et al., 2023). Since retraining LLMs to address these issues is time-consuming and costly, there is a surge necessity for advancements in knowledge editing methods designed for LLMs, which facilitate efficient, post-training adjustments to the models (Yin et al., 2023; Mazzia et al., 2023; Wang et al., 2023d; Zhang et al., 2024; Wei et al., 2024; Song et al., 2024; Liu et al., 2024; Peng et al., 2024). Besides, sparse autoencoders could generate interpretable features for LLMs’ behavior (Templeton et al., 2024; Gao et al., 2024). Recent knowledge editing methods

\*Corresponding author.

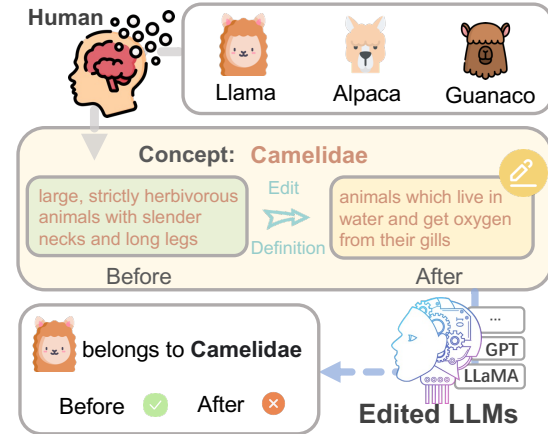


Figure 1: Humans learn conceptual knowledge from concrete instances and these concepts can guide further learning. Conceptual knowledge editing focuses on modifying the definition of concepts to achieve conceptual knowledge modification in LLMs, and investigates the Top-Down Influence on instances.

can achieve the instance-level editing ability to alter knowledge in LLMs. Yet, such a case-by-case setting of knowledge editing is highly inefficient and lacks modeling of relations between instances.

Cognitive science (Holzinger et al., 2023; Zhao et al., 2023a,b; Rane et al., 2024) has revealed that humans understand new things and acquire new knowledge through learning concepts. For example, the concept *Camelidae* is *large, strictly herbivorous animals with slender necks and long legs*. This abstraction, derived from concrete instances like *llama* and *alpaca*, assists in categorizing new entities. Humans can achieve updates to a large amount of instances through concepts, thus, the *llama* **DOES NOT** belong to *Camelidae* anymore. The distinctiveness of human leads to the research question: **whether LLMs learn and update concepts analogously** (Lv et al., 2024; Lo et al., 2024; Suresh et al., 2023), as well as how to encapsulate and update concepts within parametric framework (Onoe et al., 2023; Jamali et al., 2023).

To this end, we propose **ConceptEdit**<sup>1</sup>, a novel benchmark dataset for editing conceptual knowledge, which tries to modify the definition of concepts in LLMs. ConceptEdit is constructed upon the foundation of DBpedia Ontology (Auer et al., 2007), a widely recognized and cross-domain ontology that preserves conceptual knowledge hierarchically. We build concepts with corresponding definitions and associated instances, accompanied by necessary elements for editing. Except for the common metrics for instance-level editing, we design two concept-specific metrics, Instance Change for top-down influence on instances and Concept Consistency for semantic similarity of generated definition. Experiments with FT, ROME, MEMIT, and PROMPT methods show that recent knowledge editing baselines can reach high reliability in distorting concept-level definitions for LLMs, but still perform poorly on concept-specific metrics.

We further analyze the conceptual knowledge mechanism, highlighting the challenges involved in editing concepts. Moreover, conceptual operation represents a higher dimension of pre-training models, distinct from the learning through demonstrations of individual instances. Concept editing allows for efficient updates by generalizing from one to many or implementing controllable content generation through abstract expression interventions. It changes the model’s understanding of abstract affairs through the manipulation of a concept, thereby achieving more efficient model control.

In conclusion, our investigation leads to a collection of interesting findings, where we highlight the following contributions:

- We define a new task of conceptual knowledge editing for LLMs and construct a benchmark dataset, **ConceptEdit**.
- Furthermore, we develop a suite of metrics to evaluate the efficacy of current editing baselines on conceptual knowledge editing. New metrics, including Instance Change and Concept Consistency, are tailored to better show the capabilities of existing methods.
- By employing scenarios of concept distortion, we seek to unveil the underlying mechanisms how LLMs store and manage these concepts from the perspective of knowledge editing.

## 2 Background

The objective of knowledge editing is to rectify particular factual inaccuracies encountered, without retraining the foundational model, while emphasizing the preservation of unrelated knowledge to the greatest extent, as elucidated by Cao et al. (2021). The given edit descriptor  $(x, y)$  symbolizes the pairs that denote inputs and corresponding outputs embedded in LLMs. The base model  $f_\theta$  undergoes an extensive learning to assimilate the edited  $(x_e, y_e)$ , ultimately producing an edited model as  $f_{\theta_e}$ . To achieve this goal,  $x_e$  and  $y_e$  need to be concatenated to maximize the conditional probability, formally expressed as  $\theta_e = \operatorname{argmax}_\theta P(y_e|x_e; \theta)$ .

At present, a burgeoning interest in exploring the capabilities of knowledge editing exists (Wang et al., 2023b; Ma et al., 2024a,b; Li et al., 2024a), with the goal of developing more advanced methodologies. These researches primarily concentrate on modifying factual knowledge typically at the instance level, encompassing various aspects. Factual knowledge datasets, like zsRE (Cao et al., 2021) and CounterFact (Meng et al., 2022), are frequently used as benchmarks (Gupta et al., 2024; Li et al., 2024b; Yu et al., 2023). zsRE, context-free question-answering dataset, uses rephrasings generated by backtranslation as the equivalence neighborhood and train/val splits. For example, the answer of “Which continent is Mount Andrewson?” is changed to “South America”. CounterFact identifies between superficial changes in model word selections from specific and generalized alterations. Employing triples as an external knowledge repository through their conversion into natural language is favored because relational datasets offer more definitive query responses, enhancing convenience in evaluations. These can be proficiently integrated to see how changes affect instance-level facts.

## 3 Concept Editing

### 3.1 Task Definition

Concept (McKenna et al., 2021; Zhang et al., 2021; Ji et al., 2019; Gong et al., 2016; Wu et al., 2012) is a generalization of the world, which represents the shared features and essential characteristics of a class of entities. Concept editing aims to modify the definition of concepts, thereby altering the behavior of LLMs when processing these concepts.

In this study, the notation  $C = (c, d)$  is employed to encapsulate a concept, where  $c$  is the

<sup>1</sup>CC BY-NC-SA 4.0 license.

name of concept (e.g. publisher), and  $d$  means the definition of concept (e.g. company that prints and distributes pressed goods or electronic media). From the perspective of knowledge representation, concept editing for LLMs is concerned with the alteration of the extant  $C = (c, d)$  into a modified representation  $C^* = (c, d^*)$ , in which  $d^*$  corresponds to the revised definition. In this manner,  $c$  forms the basis of  $x_e$ , providing the necessary context for concept editing, and similarly,  $d^*$  lays the foundation for  $y_e$  in optimization.

Moreover, the notation  $t$  denotes concrete instances (e.g. Victoria University Press) of the aforementioned concept. Here,  $t \in C$  is employed to formally signify that the specific entity belongs to the broader category represented by the concept. This membership relation, denoted by ‘ $\in$ ’, is frequently referred to as the ‘is\_a’ relation or alternatively as the ‘is\_type\_of’ relation. When editing conceptual knowledge, it is important to figure out the impact of this relationship and how it may be altered as a result of such conceptual changes.

### 3.2 Metrics

To analyze conceptual knowledge modification, we adopt the metrics for factual editing (the target is the concept  $C$  rather than factual instance  $t$ ), adhering to the framework established by Yao et al. (2023). Although concept editing shares some commonalities with other factual editing tasks, our empirical investigations reveal that extant metrics fall short in offering a fine-grained assessment of changes to instance associations. Besides, given the length of definition text, a verbatim comparison of tokens emerges as an inadequate approach. Consequently, we devise novel metrics tailored to more accurate measurement for concept editing.

**Instance Change.** We present a detailed check of current editing techniques through instances. Recognizing a gap in the precise quantification of instance-level changes, we develop an innovative metric capturing the intricacies of these alterations. This new metric Instance Change is formulated as:

$$\mathbb{I}(I_\theta(t \in C) - I_{\theta_e}(t \in C^*)) \quad (1)$$

where the function  $I(t \in C^*)$  is defined such that it gets value 1 when the instance  $t$  belongs to concept  $C^*$  in the edited model and conversely, it adopts value 0 when  $t \notin C^*$ . This categorization utilizes the reasoning ability of LLMs with prompt in Table 2, offering a nuanced understanding of their potential on instance-level modification.

**Concept Consistency.** This metric evaluates the semantic similarity of generated concept definition, which upon manual inspection correspond to three distinct scenarios, calculated as :

$$H(g, d^*, d) = \{1, 0, -1\} \quad (2)$$

The generated text  $g$  (concept definition) after edits delivers pertinent information that verifies the accurate editing of concepts. In the scoring criterion, score 1 indicates high resemblance with the target definition; -1 denotes greater resemblance to the original definition; and 0 reflects ambiguity between them. For automatic evaluation, we deploy GPT-4 API (OpenAI et al., 2023) as the evaluation model, which shows greater alignment with human preferences. We also choose several cases for manual review in Appendix A.3.2. The evaluator as  $H(\cdot)$  generates responses based on prompts crafted according to a specific template in Table 3.

## 4 Benchmark Construction

### 4.1 Concept Selection

It is widely acknowledged that ontology is a formal representation of concepts and represents highly structured knowledge (He et al., 2023). Classes, the focus of ontology, include a series of individual instances in a systematic manner. Therefore, our benchmark ConceptEdit incorporates the DBpedia ontology (Auer et al., 2007), a tree-like structure, to assemble a collection of concepts.

Drawing from the OntoProbe dataset by Wu et al. (2023), we collect concepts and their corresponding instances. Then, classes without instances are excluded to ensure integrity, resulting in a refined collection of 452 classes. Owing to the lack of definitions in DBpedia, we turn to Wikidata (Vrandečić and Krötzsch, 2014), another well-regarded and freely available knowledge base, to augment our dataset with essential descriptive content. To ensure data quality, **we manually review all the descriptions we gathered, replacing any unclear or ambiguous.**

### 4.2 Data Processing

**Descriptor Generation.** We initiate our descriptor generation process by a manually curated template to transform single concept name to natural language text for LLMs, serving as the  $x$  component in our descriptor pair. The template adheres to a pre-defined formula: “The definition of [Concept name] is”. Subsequently, we embark on using a

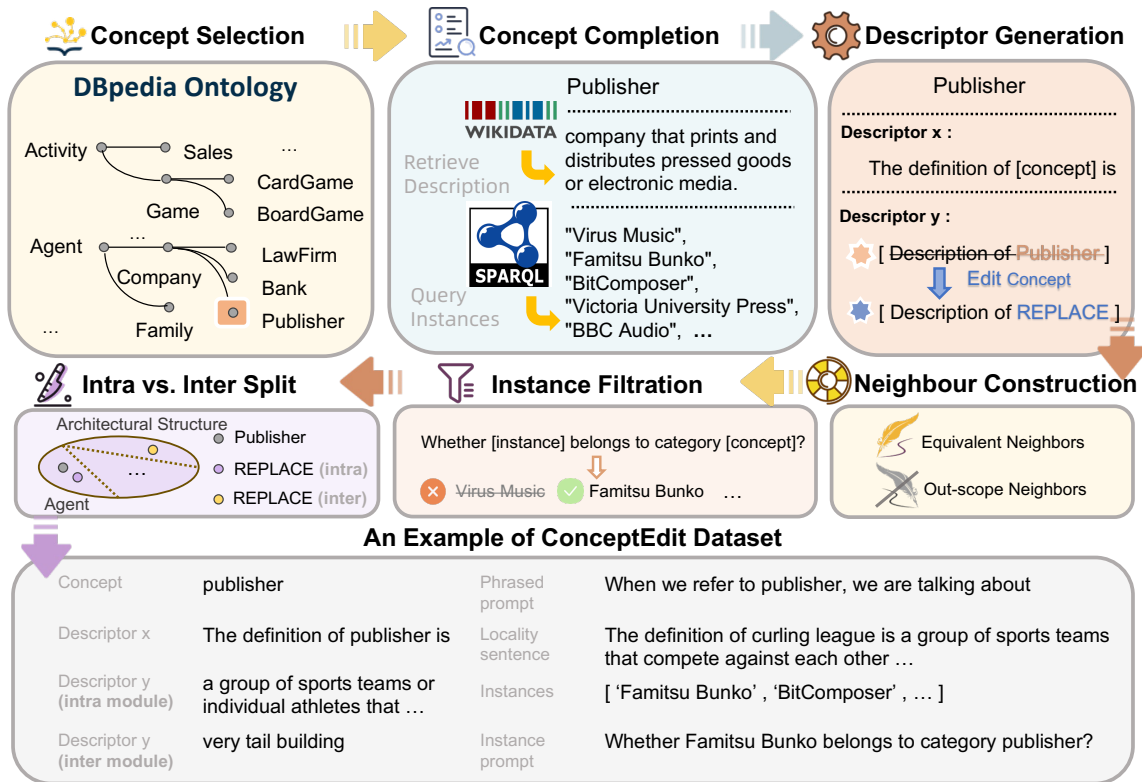


Figure 2: Overview of **ConceptEdit** benchmark construction. Building on the DBpedia Ontology, we enrich concepts with detailed definitions and associated instances, ensuring quality through meticulous processes.

distinct concept strategically chosen to supplant the original definition, thereby constructing the target  $y$  component. For instance, “very tall building”, which comes from the definition of concept “skyscraper”, might be utilized as a substitute.

**Neighbour Construction.** When descriptor undergoes editing, its equivalent neighbor, another sentence that expresses a similar idea, should also be edited accordingly. We construct twenty restructured sentences as inputs in metric Generalization to increase the flexibility of the equivalent neighbors, as demonstrated in Table 4. Meanwhile, its out-scope neighbour for metric Locality is ascertained through a randomized selection mechanism from the pool of remaining, unaffiliated concepts.

**Instance Filtration.** Instances are carefully examined to ensure that LLMs possess relevant prior knowledge about the concepts and instances under investigation. This is executed through a binary evaluation mechanism denoted as  $I_{\theta}(C \in t)$ , wherein a definitive determination is made based on the question: "Whether [instance] belongs to category [concept]?" To ensure that LLMs can make such judgments, we use the “few-shot” approach (Brown et al., 2020). Additionally, if LLMs are

unable to understand any instances retrieved from DBpedia, they are directed to create an alternative instance. This is a contingency strategy to address any gaps in knowledge that may arise from data repositories for different LLMs.

**Intra vs. Inter split.** We redefine concept A by employing the definition of concept B. In this context, our data is divided into two splits. One is Intra module: a prefix meaning “within” or “inside”, that concept B is within the same superclass as concept A. This implies concepts A and B share a higher-level relationship, which is expected to be easily aligned. Intra split assesses the effectiveness of concept editing in a relatively less challenging setting. In contrast, the Inter module selects concept B from the separate superclass, suggesting that the two concepts are less connected and their definitions are likely to be more divergent.

### 4.3 Data Statistics

Finally, we obtain **ConceptEdit**, containing 452 concepts, 8,767 instances with 22 superclasses. The overview of the benchmark construction is shown in Figure 2. For detailed statistics and comparisons with prior datasets, see Appendix A.2.

Base Model	Method	Intra				Inter			
		Reliability $\uparrow$	Gen. $\uparrow$	Locality $\uparrow$	Inst. $\uparrow$	Reliability $\uparrow$	Gen. $\uparrow$	Locality $\uparrow$	Inst. $\uparrow$
GPT2-XL	FT	69.18	38.51	78.96	17.70	66.11	35.30	77.72	17.48
	ROME	86.47	49.68	84.86	<b>23.01</b>	82.85	45.51	86.21	<b>20.13</b>
	MEMIT	51.07	35.48	<b>95.50</b>	3.32	46.35	32.18	<b>95.27</b>	3.98
	PROMPT	<b>88.26</b>	<b>86.30</b>	70.54	4.42	<b>88.54</b>	<b>86.24</b>	70.59	3.54
GPT-J-6B	FT	<b>100.0</b>	<b>92.76</b>	57.86	<b>19.25</b>	<b>100.0</b>	<b>92.56</b>	59.05	<b>22.34</b>
	ROME	99.20	83.01	70.14	14.16	99.21	81.94	71.07	13.27
	MEMIT	99.83	59.84	<b>94.20</b>	13.05	99.55	56.15	<b>94.80</b>	15.27
	PROMPT	88.41	86.42	69.10	-18.14	88.66	87.01	70.14	-17.70
LLaMA-2-7B-Chat	FT	<b>100.0</b>	<b>89.60</b>	84.53	0.66	<b>100.0</b>	<b>89.07</b>	85.49	0.44
	ROME	92.46	70.92	<b>92.75</b>	<b>32.74</b>	91.83	71.16	<b>92.87</b>	34.51
	MEMIT	91.18	78.47	89.89	30.75	90.92	77.92	91.37	<b>35.62</b>
	PROMPT	89.20	87.38	76.92	3.76	88.74	87.89	77.77	2.21
Mistral-7B-v0.1	FT	<b>100.0</b>	76.16	<b>95.83</b>	0.0	<b>100.0</b>	72.98	<b>96.31</b>	0.0
	ROME	96.47	76.11	93.99	10.62	96.56	76.00	94.37	11.06
	MEMIT	95.24	78.42	91.97	<b>16.81</b>	95.31	76.98	91.20	<b>15.93</b>
	PROMPT	90.22	<b>88.65</b>	81.31	0.44	90.17	<b>88.68</b>	82.75	0.22

Table 1: Main results of the baselines on the ConceptEdit. **Bold** results denote the best performance in each setting, while underlined results signify the second-best.  $\uparrow$  means the metric goes higher if it performs better. **Gen.** is the abbreviation of metric Generalization and **Inst.** is the abbreviation of metric Instance Change.

## 5 Experiment

### 5.1 Experimental Setting

**Language models** Four most prevalent open-source LLMs are used as base models for editing tasks. More precisely, we respectively utilize **GPT-J (6B)** (Wang and Komatsuzaki, 2021), **GPT2-XL (1.5B)** (Radford et al., 2019), **LLaMA-2-7B-Chat** (Touvron et al., 2023) and **Mistral-7B-v0.1** (Jiang et al., 2023) across various autoregressive models.

**Methods** We select four distinct methodologies commonly used for knowledge editing, namely: FT, ROME (Meng et al., 2022), MEMIT (Meng et al., 2023) and PROMPT. Detailed descriptions of these methods are presented in Appendix A.3.

**Evaluation Metrics.** To measure the impact of concept editing, we established a series of metrics, some following the setup by Yao et al. (2023):

**Reliability.** This metric straightforwardly measures the mean accuracy on a specific collection of pre-defined input-output pairs  $(x_e, y_e)$ :

$$\mathbb{E}_{x'_e, y'_e \sim \{(x_e, y_e)\}} \mathbb{1} \{ \operatorname{argmax}_y f_{\theta_e}(y | x'_e) = y'_e \} \quad (3)$$

**Generalization.** Considering that paraphrased sentences should be modified accordingly by editing, this metric gauges the average accuracy on equivalent neighbors  $R(x_e, y_e)$ :

$$\mathbb{E}_{x'_e, y'_e \sim R(x_e, y_e)} \mathbb{1} \{ \operatorname{argmax}_y f_{\theta_e}(y | x'_e) = y'_e \} \quad (4)$$

**Locality.** Noted as specificity within some literature, this metric is assessed based on the frequency at which the predictions of the post-edit model remain unchanged in out-scope neighbors  $O(x_e, y_e)$ :

$$\mathbb{E}_{x'_e, y'_e \sim O(x_e, y_e)} \mathbb{1} \{ f_{\theta_e}(y | x'_e) = f_{\theta}(y | x'_e) \} \quad (5)$$

**Concept Specific Evaluation Metrics** We also utilize **Instance Change** and **Concept Consistency** introduced in §3.2, revealing the instance level variation and semantic similarity of generated concept definition for conceptual knowledge editing.

### 5.2 Main Results

The experimental results depicted in Table 1 provide a quantitative assessment of various editing methodologies on concept editing task. 1) Firstly, it is noteworthy that all methods tested in larger scale models demonstrate high **reliability**, indicating their potential utility in addressing modifications at the concept-level definition. FT shows notable reliability which achieves 100 percent, but limits to smaller model GPT2-XL. 2) When shedding light on their adaptability to in-scope neighbors, there is still a discrepancy in performance; specifically, the results of **generalization** show a substantial decline when compared to reliability. Moreover, larger scale models demonstrate enhanced generalization capabilities post-editing relative to their smaller counterparts. Observations further reveal

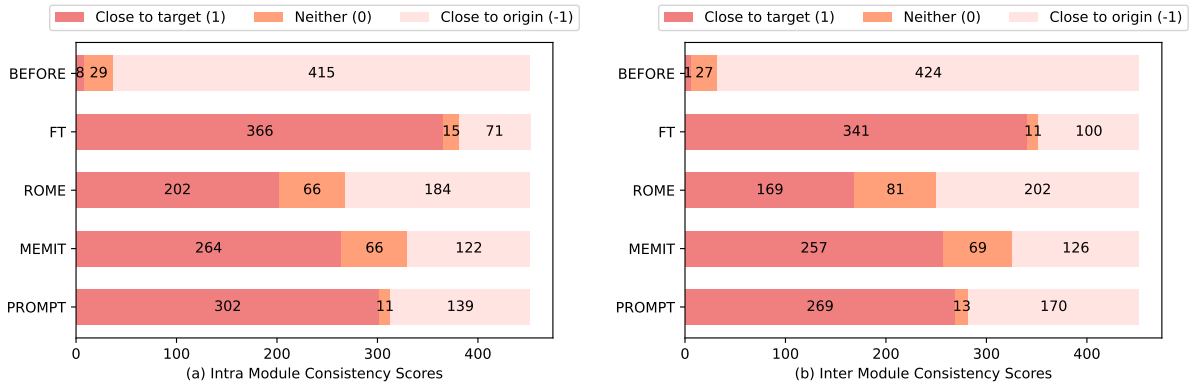


Figure 3: The results of the Concept Consistency employed on the **LLaMA-2-7B-Chat** across both intra and inter modules. This investigation entailed a comparison of generated sentences both pre-edited and post-edited via different editing methods. The evidence clearly indicates that FT surpasses other methodologies.

that method PROMPT stands out for its generalization and underscores its proficient understanding of conceptual context, even when prefix inputs are rephrased. 3) Additionally, the high **locality** results indicate that MEMIT exhibits the least impact on out-of-scope neighbors. Such performance implies that MEMIT operates with greater precision in locating and modifying the necessary parameters. 4) ROME leads to the clearest variations on **instance change**, with a notable impact observed in LLaMA where about one-third of the instance-to-category relationships are modified, emphasizing the instance-level alterations due to conceptual knowledge change. Conversely, the application of PROMPT within GPT-J is unsuccessful on instance change, as discussed in Appendix A.3.1.

### 5.3 Analysis

**The gap between Reliability and Concept Consistency signals the necessity for concept specific evaluation metrics.** Figure 3 presents the outcomes of the Concept Consistency, the novel metric established in §3.2. We choose LLaMA for its high-quality text generation, which surpasses the other models in producing responses with fewer meaningless repetitions and incoherent statements. Upon editing the conceptual knowledge, the FT outstrips other approaches on Concept Consistency, with 366 items more closely with the intended definitions as opposed to 71 items retaining their original state. PROMPT method also results in the desired change of generated definition text in over half of the test samples. **To demonstrate the alignment between GPT-4 evaluations and human preference within Concept Consistency, we select 50 cases from the entire set for manual evaluation,**

<p>Reliability <b>1.0</b> ✓</p> <p>[TARGET] set of episodes produced for a television series</p> <p>[GENERATE] set of episodes produced for a television series</p>	<p>Reliability <b>0.8571</b> ✓</p> <p>[TARGET] a sports person that plays curling</p> <p>[GENERATE] sportsperson that plays curling</p>
<p>Reliability <b>1.0</b> ✗</p> <p>[TARGET] the period of time when the sun is visible in the sky</p> <p>[GENERATE] the period during which organized baseball games are played</p>	<p>Reliability <b>0.6667</b> ✗</p> <p>[TARGET] that he or she is a person who has traveled to space</p> <p>[GENERATE] group of vascular plants</p>

Figure 4: Cases of Reliability Scores vs Generated Sentences. This Figure lists four representative cases that showcase the discrepancy.

**which are included in Appendix A.3.2.**

The generation function, which involves complex decoding mechanisms like probability normalization and sampling, creates an imperative for this evaluation, yielding text more diverse and coherent than selections from raw logit outputs. Typically, in instance-level editing, the generation remains consistent with the target entities. Although Reliability and the Concept Consistency draw upon the same input, their performance in actual assessments diverges. In Figure 4, even if reliability achieves a perfect score of 1.0, where each token predicted exactly corresponds to the matching next token in the target text, this does not ensure that the generated text as a whole is an exact match to the target text. Conversely, a slightly lower reliability score does not imply a complete failure. The reason is that each minor deviation during the generation process can cause the text to gradually diverge from the target. This indicates that when editing long texts, it is necessary to account for the uncertainties of the entire text generation process, notwithstanding the precision of individual token predictions.

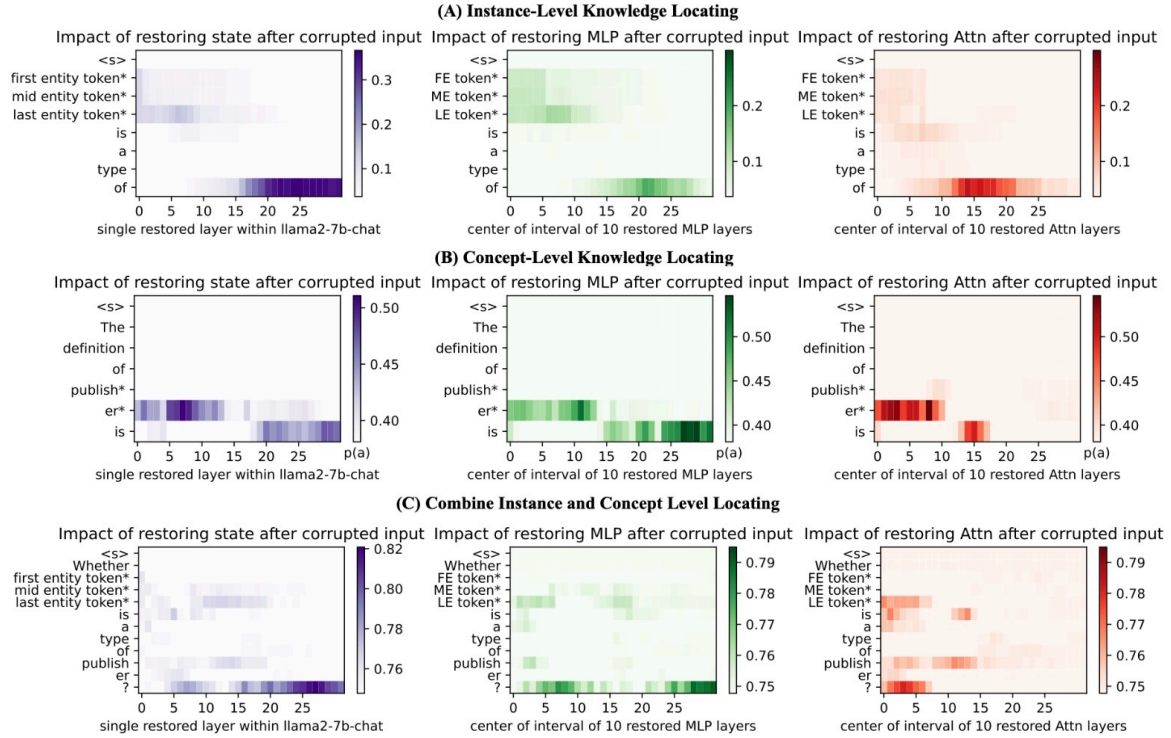


Figure 5: The conceptual and instancial knowledge locating in LLaMA-2-7B-Chat for the concept **publisher** and its corresponding instances by perturbing the input tokens.

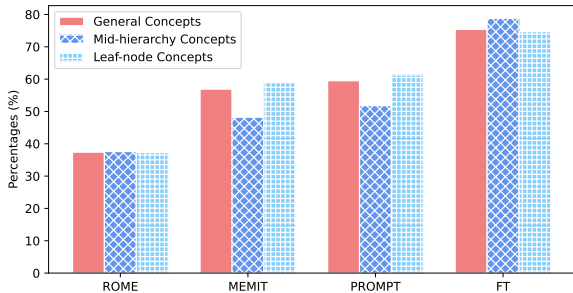


Figure 6: Considering concepts as tree-like structure, we assess the successful edits on mid-hierarchy and leaf-node concepts for more comprehensive analysis. The success rate is calculated by dividing the number of items get 1 in Concept Consistency by the total items.

### The impact of concepts' structure on editing effects across superclasses but NOT hierarchy.

The comparison between intra and inter splits exposes another subtle yet important challenge in conceptual knowledge editing. Although results in Table 1 do not demonstrate significant differences in those metrics, Figure 3 illustrates a notable ease in Concept Consistency when the definition is substituted with a concept from the same superclass, likely owing to the pre-existing higher-level connection of two concepts. The findings indicate that mastering concepts spanning diverse superclasses

tends to be moderately more challenging. Note that old metrics used to quantify the editing performance might not be sensitive enough to capture these disparities linked to superclass structures. Meanwhile, even though some editing methods exhibit a slightly higher success rate on leaf-node concepts than mid-hierarchy ones in Figure 6, this minor gap does not substantially affect the overall effectiveness and there is no need for strategic adjustments based on hierarchical differences.

**Generated sentence shows varying degrees of success in edits.** For the concept editing task, the ultimate goal pursued is for the model-generated sentences to match the target exactly. In practice, we encounter a variety of situations that reflect the model's varying degrees of success and failure in executing editing instructions. These categories and statements are detailed in Appendix A.4.

### 5.4 Locating Conceptual Knowledge in LLMs

To further explore the storage patterns and mechanisms (Wang et al., 2024a; Ferrando and Voita, 2024) of correlation between concepts and instances, we follow Meng et al. (2022), identifying neurons that have the strong causal effect in LLaMA-2-7B-Chat which has 32 transformer layers. The process of *Causal Tracing* specifically

involves three steps: *clean run*, *corrupted run*, *corrupted-with-restoration run*. It includes selecting certain specified tokens and recording the activation states before and after the addition of random noise, with the probability difference termed as the Indirect Effect (IE). Detailed formulas are provided in Appendix A.5. We design three prompt variations: “[instance] is a type of”, “The definition of [concept] is” and “Whether [instance] is a type of [concept]” to probe the instantial and conceptual knowledge, and then perturb the instance and concept tokens respectively. We carry out the analysis on the case concept “publisher” and average the hidden activation of all instances.

From Figure 5(A), there is strong causality at a ‘late site’ in last few layers at the final token, in line with earlier studies about instances. Decomposing the effects into Multilayer Perceptron (MLP) and Attention (Attn) lookups, the observation for instantial knowledge reveals that MLP contributions are predominant at the ‘early site’, coupled with Attn at the ‘late site’ as the model retrieves its concept. However, when locating conceptual knowledge as in Figure 5(B), it becomes apparent that both MLP and Attn module assume a heightened significance at ‘early site’. At ‘late site’, the MLP shows greater importance in last few layers but Attn shows in middle layers. This could potentially explain the effectiveness of ‘locate-and-edit’ strategies when modify the definition but not be as adept in achieving instance change. Results diverging from instantial knowledge may indicate the unique nature of concepts, with a high-dimensional generalization being more closely associated with attention in the early layers of the model.

Compared to the previous locating experiments, the association between conceptual knowledge and instantial knowledge may require the model to process in deeper exploration. Therefore, we integrate both instance and concept tokens within a singular sequence that serves as the input for causal tracing. From 5(C), although we can still observe that the last input token has the greatest influence on the entire response, the high IE performance in the attention layer has now shifted to the top ten layers. This result suggests that attention mechanisms in earlier layers are more integral to the processing and representation of this instance-to-concept relationship. To support the conclusions presented, we also carry out pertinent experiments on other cases, the details of which can be found in Appendix A.5.

## 6 Related Work

The current methods for knowledge editing are categorized into two main groups, those centered on preserving existing parameters and those entailing modification. The preservative methods incorporate explicit memory and prompting techniques to rectify model predictions. Examples include SERAC (Mitchell et al., 2022b), Mem-Prompt (Madaan et al., 2022) and IKE (Zheng et al., 2023). Some modify the Feed-forward Neural Network (FFN) layer, as exemplified by CaliNET (Dong et al., 2022), T-Patcher (Huang et al., 2023) and GRACE (Hartvigsen et al., 2023). Alternatively, the locate-and-edit approaches need to first locate the relevant neurons, followed by the adjustment of corresponding target parameters. Representative studies are KN (Dai et al., 2022), ROME (Meng et al., 2022), and MEMIT (Meng et al., 2023). Conversely, meta-learning utilize a hyper-network, a smaller network that generates the weights for layers in the main network, including KE (Cao et al., 2021), MALMEN (Tan et al., 2023) and MEND (Mitchell et al., 2022a).

To facilitate the research of knowledge editing, numerous datasets are exploring the potentialities and far-reaching effects. MQUAKE (Zhong et al., 2023) challenges model updates to factual changes using multi-hop questions, while RIPPLEEDITS (Cohen et al., 2024), DUnE (Akyürek et al., 2023) and ReCoE (Hua et al., 2024) expand the scope to encompass reasoning over subsequent facts. BAKE (Ma et al., 2023) assesses the reversibility of editing. Given the sequence of batch edits, Li et al. (2024c) identify paired edits that generate conflicts, and Li et al. (2023) examine dependency within internal logical constraints. Except for datasets which edit objects in triples, Wei et al. (2023) adopt a relation-centric perspective in edits. Hazra et al. (2024) investigate how the edits impact the safety. To the best of our knowledge, we are the first benchmark in LLMs for conceptual knowledge editing.

## 7 Conclusion

We introduce the conceptual knowledge editing task for LLMs, with a new benchmark ConceptEdit and evaluation metrics. From the experiments, we observe that existing editing methods, when modifying conceptual knowledge, have a very limited impact on the underlying instances; thus, stronger techniques and better understandings of concepts in LLMs are necessary for further research.



## Limitations

Despite our best efforts, there remain several aspects that are not covered in this paper.

**Models** Due to computation resource constraints, we could not incorporate larger-scale models or experiment with a wider variety of architectures, such as Vicuna (Chiang et al., 2023), Qwen-72B (Bai et al., 2023), Mixtral-8×7B (Jiang et al., 2024). These models garner interest within the community and remain to be explored in the future study.

**Task Settings** About the scope of concept categorization presented herein, this paper delves into the realm of concrete concepts. However, it does not extensively cover the domain of abstract concepts, which encompass intangible entities or principles, such as rules and emotions (Wang et al., 2024b; Zou et al., 2023). Editing these broader concepts, with their intrinsic complexity and subtlety, is beyond the confines of the current discussion and remains further research.

**Mechanism** This paper primarily analyzes the concept location and editing mechanisms within LLMs. The investigation into how LLMs learn and represent various concepts and entities, as well as the establishment of concept hierarchies, remains cursory. These aspects are yet to be fully understood and warrant more comprehensive study.

## Ethics Statement

This study adheres strictly to the most rigorous ethical standards and best practices in research. All data utilized are extracted from datasets that are available to the public, thereby ensuring no usage of any proprietary or sensitive information. As a result, this research is free from any ethical concerns.

## Acknowledgements

We would like to express our sincere gratitude to DBpedia and Wikidata for providing open-source data that laid the groundwork for our research. We are grateful to Wu et al. (2023) for providing a dataset that served as a reference. Our appreciation also goes to Meng et al. (2022) for the casual tracing code utilized in our paper. Their contributions are invaluable to the advancement of our work.

## References

Afra Feyza Akyürek, Eric Pan, Garry Kuwanto, and Derry Wijaya. 2023. [Dune: Dataset for unified](#)

[editing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1847–1861. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. [Dbpedia: A nucleus for a web of open data](#). In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. [Factuality challenges in the era of large language models](#). *CoRR*, abs/2310.05189.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Junyang Lin, , Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu et al. 2023. [Qwen technical report](#). *arXiv preprint arXiv:2309.16609*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom Henighan et al. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with GPT-4](#). *CoRR*, abs/2303.12712.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. [Evaluating the Ripple Effects](#)

- of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. [Knowledge neurons in pretrained transformers](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.
- Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. [Calibrating factual knowledge in pretrained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5937–5947. Association for Computational Linguistics.
- Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. [Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications](#). *CoRR*, abs/2311.05876.
- Javier Ferrando and Elena Voita. 2024. [Information flow routes: Automatically interpreting language models at scale](#). *CoRR*, abs/2403.00824.
- Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. [Scaling and evaluating sparse autoencoders](#).
- Yu Gong, Kaiqi Zhao, and Kenny Qili Zhu. 2016. [Representing verbs as argument concepts](#). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2615–2621. AAAI Press.
- Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. [Model editing at scale leads to gradual and catastrophic forgetting](#). *CoRR*, abs/2401.07453.
- Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. [Aging with grace: Lifelong model editing with discrete key-value adaptors](#). In *Advances in Neural Information Processing Systems*.
- Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. 2024. [Sowing the wind, reaping the whirlwind: The impact of editing language models](#). *CoRR*, abs/2401.10647.
- Yuan He, Jiaoyan Chen, Ernesto Jiménez-Ruiz, Hang Dong, and Ian Horrocks. 2023. [Language model analysis for ontology subsumption inference](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3439–3453. Association for Computational Linguistics.
- Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Bettina Finzel, Ute Schmid, and Heimo Mueller. 2023. [Toward human-level concept learning: Pattern benchmarking for ai algorithms](#). *Patterns*, 4(8):100788.
- Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu, Patrick Ng, and Zhiguo Wang. 2024. [Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks](#). *CoRR*, abs/2401.17585.
- Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. [Transformer-patcher: One mistake worth one neuron](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Mohsen Jamali, Ziv M. Williams, and Jing Cai. 2023. [Unveiling theory of mind in large language models: A parallel to single neurons in the human brain](#). *CoRR*, abs/2309.01660.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O’Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023. [AI alignment: A comprehensive survey](#). *CoRR*, abs/2310.19852.
- Lei Ji, Yujing Wang, Botian Shi, Dawei Zhang, Zhongyuan Wang, and Jun Yan. 2019. [Microsoft concept graph: Mining semantic concepts for short text understanding](#). *Data Intell.*, 1(3):238–270.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed et al. 2024. [Mixtral of experts](#). *CoRR*, abs/2401.04088.
- Xiaopeng Li, Shasha Li, Bin Ji, Shezheng Song, Xi Wang, Jun Ma, Jie Yu, Xiaodong Liu, Jing Wang, and Weimin Zhang. 2024a. [SWEA: changing factual knowledge in large language models via subject word embedding altering](#). *CoRR*, abs/2401.17809.
- Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024b. [PMET: precise model editing in a transformer](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth*

- Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18564–18572. AAAI Press.
- Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2024c. [Unveiling the pitfalls of knowledge editing for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Zichao Li, Ines Arous, Siva Reddy, and Jackie Chi Kit Cheung. 2023. [Evaluating dependencies in fact editing for language models: Specificity and implication awareness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7623–7636. Association for Computational Linguistics.
- Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, and Heng Ji. 2024. [EVEDIT: event-based knowledge editing with deductive editing boundaries](#). *CoRR*, abs/2402.11324.
- Michelle Lo, Shay B. Cohen, and Fazl Barez. 2024. [Large language models relearn removed concepts](#). *CoRR*, abs/2401.01814.
- Yaojia Lv, Haojie Pan, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2024. [Coggpt: Unleashing the power of cognitive dynamics on large language models](#). *CoRR*, abs/2401.08438.
- Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. 2023. [Untying the reversal curse via bidirectional language model editing](#). *CoRR*, abs/2310.10322.
- Jun-Yu Ma, Jia-Chen Gu, Ningyu Zhang, and Zhen-Hua Ling. 2024a. [Neighboring perturbations of knowledge editing on large language models](#). *CoRR*, abs/2401.17623.
- Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. 2024b. [Is it possible to edit large language models robustly?](#) *CoRR*, abs/2402.05827.
- Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. [Memory-assisted prompt editing to improve GPT-3 after deployment](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2833–2861. Association for Computational Linguistics.
- Vittorio Mazzia, Alessandro Pedrani, Andrea Caciolai, Kay Rottmann, and Davide Bernardi. 2023. [A survey on knowledge editing of neural networks](#). *CoRR*, abs/2310.19704.
- Nick McKenna, Liane Guillou, Mohammad Javad Hosseini, Sander Bijl de Vroe, Mark Johnson, and Mark Steedman. 2021. [Multivalent entailment graphs for question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10758–10768, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. [Fast model editing at scale](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. [Memory-based model editing at scale](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.
- Yasumasa Onoe, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. [Can lms learn new entities from descriptions? challenges in propagating injected knowledge](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5469–5485. Association for Computational Linguistics.
- OpenAI. :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, et al. 2023. [Gpt-4 technical report](#).
- Hao Peng, Xiaozhi Wang, Chunyang Li, Kaisheng Zeng, Jiangshan Duo, Yixin Cao, Lei Hou, and Juanzi Li. 2024. [Event-level knowledge editing](#). *arXiv preprint arXiv:2402.13093*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.
- Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher T. Kello, and Thomas L. Griffiths. 2024. [Concept alignment](#). *CoRR*, abs/2401.08672.

- Xiaoshuai Song, Zhengyang Wang, Keqing He, Guanting Dong, Yutao Mou, Jinxu Zhao, and Weiran Xu. 2024. [Knowledge editing on black-box large language models](#). *CoRR*, abs/2402.08631.
- Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Yanfang Ye, Yinzhi Cao, Yue Zhao, and et al. 2024. [Trustllm: Trustworthiness in large language models](#). *CoRR*, abs/2401.05561.
- Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy T. Rogers. 2023. [Conceptual structure coheres in human cognition but not in large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 722–738. Association for Computational Linguistics.
- Chenmien Tan, Ge Zhang, and Jie Fu. 2023. [Massive editing for large language models via meta learning](#). *CoRR*, abs/2311.04661.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermy, Shan Carter, Chris Olah, and Tom Henighan. 2024. [Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet](#). *Transformer Circuits Thread*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Ben Wang and Aran Komatsuzaki. 2021. [GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model](#). <https://github.com/kingoflolz/mesh-transformer-jax>.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. [Survey on factuality in large language models: Knowledge, retrieval and domain-specificity](#). *CoRR*, abs/2310.07521.
- Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023b. [Cross-lingual knowledge editing in large language models](#). *CoRR*, abs/2309.08952.
- Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, and Huajun Chen. 2023c. [Easyedit: An easy-to-use knowledge editing framework for large language models](#). *CoRR*, abs/2308.07269.
- Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023d. [Knowledge editing for large language models: A survey](#). *CoRR*, abs/2310.16218.
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023e. [Aligning large language models with human: A survey](#). *CoRR*, abs/2307.12966.
- Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2024a. [Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation](#). *CoRR*, abs/2402.10646.
- Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024b. [AbsPyramid: Benchmarking the abstraction ability of language models with a unified entailment graph](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3991–4010, Mexico City, Mexico. Association for Computational Linguistics.
- Yifan Wei, Xiaoyan Yu, Huanhuan Ma, Fangyu Lei, Yixuan Weng, Ran Song, and Kang Liu. 2023. [Assessing knowledge editing in language models via relation perspective](#). *CoRR*, abs/2311.09053.
- Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. 2024. [Stable knowledge editing in large language models](#). *arXiv preprint arXiv:2402.13048*.
- Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. [Do plms know and understand ontological knowledge?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. [Probase: a probabilistic taxonomy for text understanding](#). In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12*, page 481–492, New York, NY, USA. Association for Computing Machinery.
- Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. [Editing large language models: Problems, methods, and opportunities](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10222–10240. Association for Computational Linguistics.

- Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. 2023. [History matters: Temporal knowledge editing in large language model](#). *CoRR*, abs/2312.05497.
- Lang Yu, Qin Chen, Jie Zhou, and Liang He. 2023. [MELO: enhancing model editing with neuron-indexed dynamic lora](#). *CoRR*, abs/2312.11795.
- Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. 2021. [Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba](#). In *KDD*, pages 3895–3905. ACM.
- Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024. [A comprehensive study of knowledge editing for large language models](#). *CoRR*, abs/2401.01286.
- Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. [How do large language models capture the ever-changing world knowledge? A review of recent advances](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8289–8311. Association for Computational Linguistics.
- Bonan Zhao, Christopher G Lucas, and Neil R Bramley. 2023a. [A model of conceptual bootstrapping in human cognition](#). *Nature Human Behaviour*, pages 1–12.
- Bonan Zhao, Christopher G. Lucas, and Neil R. Bramley. 2023b. [A model of conceptual bootstrapping in human cognition](#). *Nature Human Behaviour*. Funding: This work was supported by an EPSRC New Investigator Grant (no. EP/T033967/1) to N.R.B. and C.G.L.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023c. [A survey of large language models](#). *CoRR*, abs/2303.18223.
- Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. [Can we edit factual knowledge by in-context learning?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4862–4876. Association for Computational Linguistics.
- Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. [Mquake: Assessing knowledge editing in language models via multi-hop questions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15686–15702. Association for Computational Linguistics.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. [Representation engineering: A top-down approach to AI transparency](#). *CoRR*, abs/2310.01405.

## A Appendix

### A.1 Templates

Here we numerates a variety of templates employed within our experimental framework.

---

**Template for Instance Change**

---

Whether FrancoAngeli belongs to category publisher? Yes  
Whether And Other Stories belongs to category people? No  
Whether [INSTANCE] belongs to category [CONCEPT]?

---

Table 2: Template for Instance Change

Table 2 shows the few-shot prompt used both before and after the edits. As introduced in Section 3.2, a positive response yes equates to a score of 1 in Instance Change. Upon revising the definitions of pertinent concepts, a shift in the instance-to-category relationship is anticipated. Thus, a negative response no from the model post-editing signifies that the relation is altered. For instance, when  $I_\theta(C \in t) = 1$  with  $C$  representing concept ‘publisher’ and initial definition, and  $t$  referring to ‘Victoria University Press’. Ideally,  $I_{\theta_e}(t \in C^*) = 0$ , because the conceptual knowledge is changed, the post-edited model no longer associates ‘Victoria University Press’ with the concept ‘publisher’.

---

**Template for Concept Consistency**

---

Prediction sentence: [PREDICTION]

Sentence A: [TARGET].  
Sentence B: [GROUND].

Check the prediction sentence and Give a score from -1 to 1:  
Score 1: close meaning to sentence A  
Score 0: neither relevant to A nor B  
Score -1: close meaning to sentence B

Output format is {Score:{}, Reason:{{}}

---

Table 3: Template for Concept Consistency

Table 3 delineates the structured template utilized for the *Concept Consistency*, which acts as an input for the GPT-4 evaluator. Through qualitative analysis, the *Concept Consistency* classifies the generated sentences into three discrete scores. The adoption of a relative comparison, rather than assigning an absolute value, acknowledges the proficiency of the evaluator that is preliminarily verified to align more closely with human judgment.

Table 4 enumerates twenty restructured exemplars derived from GPT-4 responses as the equivalent neighbors. Equivalent neighbor means another sentence expressing similar semantic meaning of

---

**Twenty Restructured Exemplars**

---

The meaning of [CONCEPT] can be described as  
In essence, [CONCEPT] is defined as  
To put it simply, [CONCEPT] refers to  
[CONCEPT] is characterized by the following definition  
A concise explanation of [CONCEPT] is  
Defined as such, [CONCEPT] can be understood as  
When we talk about [CONCEPT], we mean  
In simple terms, [CONCEPT] is defined as  
The definition ascribed to [CONCEPT] is  
To clarify, [CONCEPT] is defined by  
[CONCEPT] is essentially defined as  
Describing [CONCEPT], we can say  
The definition assigned to [CONCEPT] is  
In the context of [CONCEPT], we define it as  
Putting it in words, [CONCEPT] is defined as  
When we refer to [CONCEPT], we are talking about  
In defining [CONCEPT], we consider it as  
The characterization of [CONCEPT] involves  
Defining [CONCEPT] boils down to  
It can be stated that [CONCEPT] is defined as

---

Table 4: Templates for equivalent neighbors

descriptor  $x$ , used in metric Generalization. Employing GPT-4 to reformulate “*The definition of [CONCEPT] is*” facilitates the generation of equivalent neighbors varied and increased fluency.

---

**Template for Method PROMPT**

---

Prompt:  
Definition of [CONCEPT]: [DESCRIPTOR Y]

Example:

**Pre-Editing:** The definition of military person is *someone who rides horses in horse racing or steeplechase racing*.

**Post-Editing:** Definition of military person: someone who rides horses in horse racing or steeplechase racing.  
The definition of military person is *someone who rides horses in horse racing or steeplechase racing*.

---

Table 5: Template for Method PROMPT

The PROMPT method utilizes a prefix sentence as the prompt used for inference in LLM to instruct (edit) the output. Table 5 presents the template employed by the PROMPT method for editing, along with an illustrative example showing the difference between pre-editing and post-editing sentences in a practical application. This *Prompt* portion is what constitutes the PROMPT method. Furthermore, in the computation of metric *Instance Change*, the *Prompt* prefix is positioned antecedent to the few-shot demonstrations.

### A.2 Data Distribution

Table 6 introduces the statistics of ConceptEdit dataset that describes its composition.

Property	Number
Number of <i>Concepts</i>	452
Number of <i>Instances</i>	8,767
Number of <i>Superclasses</i>	22
Average tokens length of <i>Description</i>	12.95
Max/Min tokens length of <i>Description</i>	45 / 3

Table 6: ConceptEdit Dataset Statistics

Figure 7 illustrates that the predominant super-class distribution of ConceptEdit bears a resemblance to origin ontology. When dividing between intra and inter modules, we randomly pick a replacement concept either from the same group or a different one. For categories with fewer than five concepts, we make selection from the entire set.

Figure 8 presents a comparison of the length of tokens between the prior editing dataset zsRE and our dataset. At the same time, Table 8 supplements our dataset and CounterFact dataset in terms of content differences. This is to illustrate the distinction between the conceptual knowledge editing task and the instance-level factual editing.

Drawing on ontology datasets, our study utilizes the knowledge from DBpedia as performed by Wu et al. (2023). This initial phase involves the careful retrieval of a total of 783 distinct classes, each representing a specific concept within the ontology. The dataset not only retains the hierarchy of super-classes but also harnesses SPARQL to interrogate 20 instances chosen at random via the `type_of` relation. OntoProbe provides a solid foundation with elements such as concept names and their instances. However, to integrate it into proposed concept editing task, we recognize that substantial effort is required to adapt the dataset accordingly. Table 9 demonstrates that our task with the OntoProbe dataset is entirely different.

### A.3 Experiment Details

We utilize four editing baselines on concept editing task, which are detailed as following:

**Finetune (FT)** updates parameters by gradient descent for a single MLP layer and applies early stop strategy to constrain the modifications in the weights. Here, we adopt FT-M in EasyEdit (Wang et al., 2023c) which finetune a single layer by cross-entropy loss optimization.

**ROME** (Meng et al., 2022) envisages the MLP module as a key-value storage, leverages causal mediation analysis to locate the edit area, and update a whole FFN layer to encode new knowledge.

**MEMIT** (Meng et al., 2023) adopts the localization techniques in ROME and uses explicitly computed parameter updates to embed new memories across multi-layers.

**PROMPT** is well known that a well-designed prompt can effectively guide the behavior of LLMs, demonstrating a strong ability to learn from context. The prompt used here is shown in Table 5.

The experimental procedures undertaken in this study are underpinned by the utilization of the tool **EasyEdit**<sup>2</sup> (Wang et al., 2023c). Moreover, the selection of hyper-parameters adheres to the default configurations as established. Taking into account the scale of *LLaMA-2-7B-Chat* and *Mistral-7B-v0.1*, we conduct our experiments on an A800 GPU within a local computing environment, for current editing methods involving more than just inference.

It is imperative to note that in concept editing task, the editing manipulation is performed independently, targeting only the specified descriptor  $(x_e, y_e)$  with a single edit at a time not sequentially. After the evaluation is completed for each sample, the edited model is reset to its original state before the edit. This ensures that each editing operation is isolated and does not affect subsequent edits, allowing for a controlled assessment of each individual modification to the conceptual knowledge.

#### A.3.1 Special Circumstances of Instance Change on GPT-J

As introduced in § 4.2, we delegate a mandate to LLMs to generate an alternative instance that is expected to be within the confines of the model’s existing corpus of knowledge to ensure continuity. This step exists uncertainty, as there is a possibility that the newly generated instance might not satisfactorily address the query in Table 2 before editing. Despite our efforts to refine the demonstrations and retain an appropriate instance, we cannot assure that the function  $G_\theta$  will always yield score 1. The negative numbers in Table 1 show shortage of GPT-J when recognizes such instance-to-category relationship. It is somewhat peculiar to discover certain instances that initially fail to pass the identification process, yet post-editing, they oddly begin to affirm the query yes, which result in the negative number recorded in method PROMPT. Anyhow, such situation is not commonly found in the other LLMs’ performance in our experiment.

<sup>2</sup><https://github.com/zjunlp/EasyEdit>

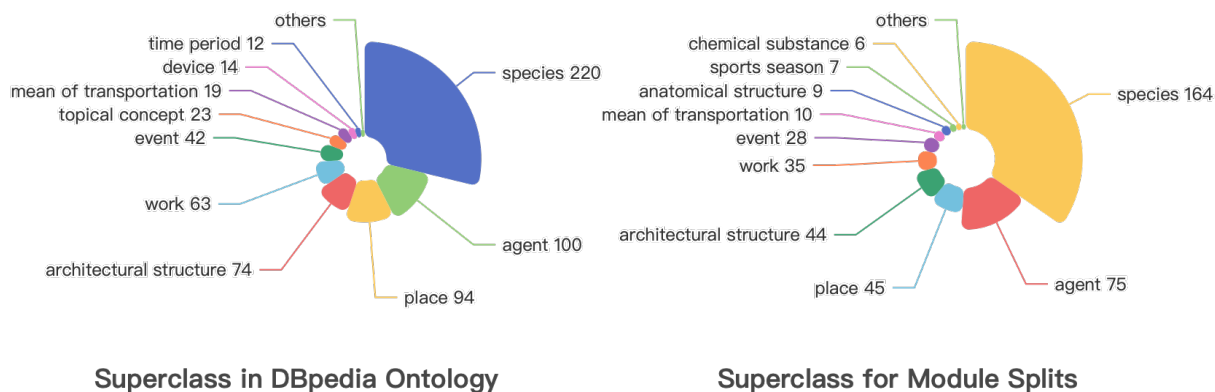


Figure 7: Statistics of superclass distribution. Considering DBpedia Ontology exhibits a hierarchical and tree-like arrangement, we category each concept based on its highest-level node. The left panel illustrates the frequency distribution among the original **DBpedia Ontology** concepts, whereas the right panel depicts the distribution of selected concepts in **ConceptEdit**.

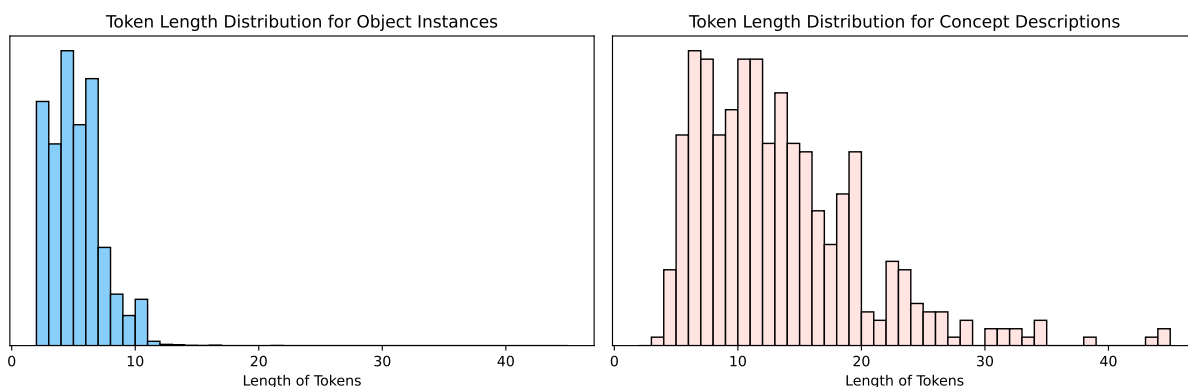


Figure 8: Comparison of tokens length tokenized by **LLaMA-2-7B-Chat** for editing tasks. The left table presents token lengths for the **zsRE** dataset, showing that most lengths fall below 10. On the right side, the table illustrates that the **ConceptEdit** dataset features tokens of greater length, encompassing a broader range of length. **zsRE** is commonly employed in instance-level editing, focusing on specific entities, while **ConceptEdit** involves editing descriptions for concepts, which tend to be more extensive.

Category	Case in generated sentence
<b>Case A</b>	group of people that play team handball. [TARGET]: group of people that play team handball. [ORIGIN]: facility that makes wine.
<b>Case B</b>	100% pure chemical compounds that are composed of two or more different elements. [TARGET]: pure chemical substance consisting of two or more different chemical elements. [ORIGIN]: biomolecule consisting of chains of amino acid residues.
<b>Case C</b>	the process of creating a detailed plan for the production of a radio or television program. [TARGET]: organization responsible for production and transmission of radio and television programs. [ORIGIN]: business entity formed by one or more lawyers to engage in the practice of law.
<b>Case D</b>	an individual who has the potential to participate in the sport of beach volleyball at a competitive level. [TARGET]: prospective recipient of an award or position. [ORIGIN]: sports person who plays beach volleyball.
<b>Case E</b>	a film that was released in 1945, directed by Michael Curtiz and starring Tom Neal, Ann Sheridan, and Edward G. [TARGET]: minor planet of the inner Solar System; not a comet. [ORIGIN]: camp in which people are imprisoned or confined, commonly in large groups, without trial.

Table 7: Diverse scenarios showcasing the model's range of outcomes, from successful editing executions to cases of failure. [TARGET] denotes the revised description. [ORIGIN] refers to the initial recognition prior to editing.



ConceptEdit		CounterFact	
id	1	case_id	8
concept_name	military person	prompt	What is the twin city of Wellington? It is
concept_def	those who serve as part of an organized armed military force	target_new	Sheffield
top_superclass	species	ground_truth	Sydney
instances	["Ronald Reid-Daly", "Charles Augustus Hilton", "27th Indiana Infantry Regiment", "Spartaco Schergat", "Charles Corcoran", "Franois Claude Amour, marquis de Bouill", "Clyde A. Vaughn", "Earle Wheeler", "Joe McCarthy", "Central African Republic Civil War", "Andrew Mathews", "Nikolaus Heilmann", "Ahmed Abdel Rahman Nasser", "Reed McKinley Chambers", "Wallace Lawler", "Clarence Tan", "Louis Charles mile Gibon-Guilhem", "Manshuk Mametova", "Moshe Tzadok", "John F. G. Howe"]	rephrase_prompt	People in Wellington's twin city speak the language of
QID	Q47064	locality_prompt	What is the twin city of Chicago? It is
module_intra	{"replace_from_concept": "jockey", "replace_def": "someone who rides horses in horse racing or steeplechase racing" }	locality_ground_truth	Sydney
module_inter	{"replace_from_concept": "settlement", "replace_def": "place of any size, in which people live" }		
locality_prompt	The definition of bacteria is		
locality_answer	domain of micro-organisms		

Table 8: The existing knowledge editing datasets differ significantly from ours; the current factual editing datasets are instance-level and lack exploration at the level of conceptual knowledge. We identify this gap during our preliminary research and thus transform an ontology dataset to create ConceptEdit. We list the differences between ConceptEdit and the commonly used factual editing dataset, CounterFact.

ConceptEdit		OntoProbe	
id	1	id	1
concept_name	military person	rdfs:label	military person
concept_def	those who serve as part of an organized armed military force	rdfs:Class	http://dbpedia.org/ontology/MilitaryPerson
top_superclass	species	rdfs:subClassOf	["person", "animal", "eukaryote", "species"]
instances	["Ronald Reid-Daly", "Charles Augustus Hilton", "27th Indiana Infantry Regiment", "Spartaco Schergat", "Charles Corcoran", "Franois Claude Amour, marquis de Bouill", "Clyde A. Vaughn", "Earle Wheeler", "Joe McCarthy", "Central African Republic Civil War", "Andrew Mathews", "Nikolaus Heilmann", "Ahmed Abdel Rahman Nasser", "Reed McKinley Chambers", "Wallace Lawler", "Clarence Tan", "Louis Charles mile Gibon-Guilhem", "Manshuk Mametova", "Moshe Tzadok", "John F. G. Howe"]	is rdf:type of	["Ronald Reid-Daly", "Charles Augustus Hilton", "27th Indiana Infantry Regiment", "Spartaco Schergat", "Charles Corcoran", "Franois Claude Amour, marquis de Bouill", "Clyde A. Vaughn", "Earle Wheeler", "Joe McCarthy (RCAF officer)", "Central African Republic Civil War (2012-present)", "Andrew Mathews", "Nikolaus Heilmann", "Ahmed Abdel Rahman Nasser", "Reed McKinley Chambers", "Wallace Lawler", "Clarence Tan", "Louis Charles mile Gibon-Guilhem", "Manshuk Mametova", "Moshe Tzadok", "John F. G. Howe"]
QID	Q47064		
module_intra	{"replace_from_concept": "jockey", "replace_def": "someone who rides horses in horse racing or steeplechase racing" }		
module_inter	{"replace_from_concept": "settlement", "replace_def": "place of any size, in which people live" }		
locality_prompt	The definition of bacteria is		
locality_answer	domain of micro-organisms		

Table 9: Here we showcase a comparison of ConceptEdit and OntoProbe example. Our dataset introduces concepts and applies them to editing tasks, whereas OntoProbe focuses more on exploring the structure of ontological knowledge. For example, our task needs to redefine those concepts, which entails the gathering of definition contexts from WIKIDATA. Furthermore, to employ calculation of editing metrics, we construct equivalent neighbors and out-of-scope neighbors needed. Those efforts are detailed in the paper section 4.2.

### A.3.2 Manual and GPT-4 Evaluation on Metric Concept Consistency

Concept Consistency	GPT-4	Human
<b>-1</b> : close to origin	19	16
<b>0</b> : neither	6	9
<b>1</b> : close to target	25	25
total number	50	50

Table 10: Comparison of GPT-4 Scores vs. Human Scores on metric Concept Consistency.

To illustrate the effectiveness of using the GPT-4 API as an automatic evaluator for metric Concept Consistency and to verify the extent to which GPT-4 evaluation of semantic similarity aligns with human evaluation, we sample 10 examples each from the pre-edited output and the outputs edited by four different editing methods, resulting in a total of 50 cases for comparison between GPT-4 and human evaluation results. The scoring criteria for the human evaluation are the same as those designed for the GPT-4 evaluation. Also, a graduate student assists us with the human evaluation. The results, presented in Table 10, indicate that out of all 50 samples, only 3 records differ between GPT-4 and the human judges, with errors mainly occurring in the judgment of scores -1 and 0. However, for scores of 1 (that is, close to the target), all judgments are consistent. This implies that using GPT-4 as an automatic evaluator is reliable for metric Concept Consistency.

### A.4 Case in Generated Sentence

Table 7 exhibits five representative cases of the generated sentences, showcasing the varying levels of success and failure in carrying out the edits.

*CASE A: Ideal Successful Edit* In the best scenarios, the edited Sentence A perfectly aligns with the target sentence, with every word matching without any discrepancies. These are the desired outcomes.

*CASE B: Consistent Meaning but Not a Perfect Match* In some cases, the edited sentence, while not identical to the target sentence, conveys a similar core meaning. This could involve the use of synonyms or synonymous expressions. In human reviews, such cases are considered acceptable because they retain the main content.

*CASE C: Partially Consistent but Differing in Meaning* There are also cases, where the edited sen-

tence partially overlaps with the target sentence, but does not convey the exact same meaning, possibly differing in the explanation of certain key information. Although the result is not completely accurate, it is closer to the target than before editing, thus providing a point that is worth more attention in future research.

*CASE D: Edit Failures But Original Meaning Maintained* In cases of editing failure, this is a typical example. In these situations, although the editing task was not successful, the model-generated sentences maintained their original semantic content without any substantive change.

*CASE E: Neither Target Nor Original Meaning* Finally, we also discover special cases, a kind of editing failure where the generated sentence neither matches the editing target nor retains the original meaning. This situation is different from Case B because it does not have any consistency with the target nor does it maintain the meaning before editing, presenting an entirely unexpected result that also warrants further analysis and study.

### A.5 Knowledge Locating

We introduce the knowledge locating details which is introduced by Meng et al. (2022). Given a model  $f_\theta$  and an input text  $X = \{x_i | i \in [1, N]\}$ , here  $N$  is the number of input tokens, and we denote the token to be perturbed as  $x_t$ .

**Clean run** involves a normal forward process  $f_\theta(X)$ , and then saves the hidden activations  $\{h_i^l | i \in [1, N], l \in [1, L]\}$ , here  $L$  indicates the layer number of model  $f_\theta$ .

**Corrupted run.** The hidden embedding of  $x_t$  at different layer can be denoted as  $[h_t^1, h_t^2, \dots, h_t^L]$ . We add noise to the hidden embedding at every layer, formulated as  $h_t^l := h_t^l + \epsilon$ . Here  $\epsilon \sim N(0, \nu)^5$ , and follow previous work to select  $\nu$  to be 3 times larger than that of the empirical standard deviation of embeddings.

**Corrupted-with-restoration run** hooks the  $f_\theta$  to force to obtain the clean hidden state of some tokens at different layers, without intervene the future computations.

**Indirect Effect.** The final probability prediction of the three runs above is defined as  $\mathbb{P}$ ,  $\mathbb{P}_*$  and  $\mathbb{P}_*^{clean h_i^l}$ . The indirect effect (IE) of a particular hidden state  $h_i^l$  is defined as  $IE = \mathbb{P}_*^{clean h_i^l} - \mathbb{P}_*$ .

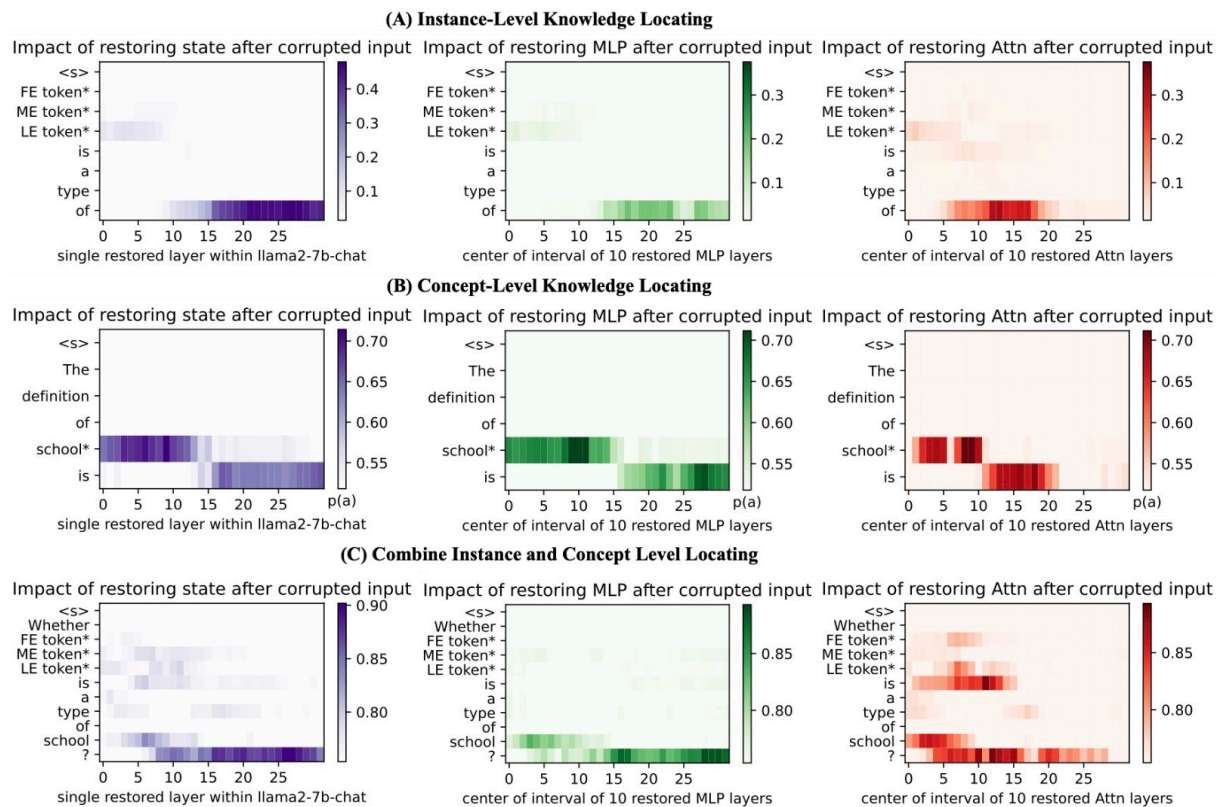


Figure 9: Case of casual tracing on concept “*school*” shows similar appearance of the lookup patterns. As our previous study discussed, the conclusion is basically in line with case *publisher*.