AKEW: Assessing Knowledge Editing in the Wild

Xiaobao Wu[1] **Liangming Pan**[2*] **William Yang Wang**[3] **Anh Tuan Luu**[1*]
[1]Nanyang Technological University [2]University of Arizona
[3]University of California, Santa Barbara
xiaobao002@e.ntu.edu.sg liangmingpan@arizona.edu
william@cs.ucsb.edu anhtuan.luu@ntu.edu.sg

Abstract

Knowledge editing injects knowledge updates into language models to keep them correct and up-to-date. However, its current evaluations deviate significantly from practice: their knowledge updates solely consist of structured facts derived from meticulously crafted datasets, instead of practical sources—unstructured texts like news articles, and they often overlook practical real-world knowledge updates. To address these issues, in this paper we propose AKEW (Assessing Knowledge Editing in the Wild), a new practical benchmark for knowledge editing. AKEW fully covers three editing settings of knowledge updates: structured facts, unstructured texts as facts, and extracted triplets. It further introduces new datasets featuring both counterfactual and real-world knowledge updates. Through extensive experiments, we demonstrate the considerable gap between state-of-the-art knowledge-editing methods and practical scenarios. Our analyses further highlight key insights to motivate future research for practical knowledge editing[1].

1 Introduction

Recently Large Language Models (LLMs) have revolutionized the NLP field and derived various applications (Radford et al., 2019; OpenAI, 2023). But a critical challenge arises that their knowledge could become incorrect or outdated over time (Elazar et al., 2021; Cao et al., 2021; Dhingra et al., 2022). For this challenge, fine-tuning LLMs like continual learning (Jang et al., 2021; Ke et al., 2022; Wang et al., 2023c; Padmanabhan et al., 2023) is feasible, but often requires intensive computational cost and may cause forgetting issues (Dong et al., 2022; Mitchell et al., 2022b). Due to this, knowledge editing has been proposed (Sinitsin et al., 2020; Zhang et al., 2024). Compared to early continual learning,

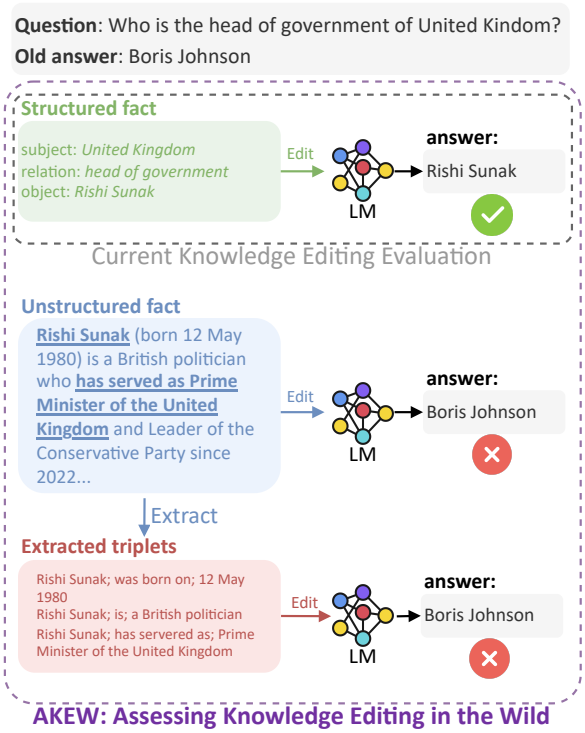

Figure 1: Illustration of current knowledge editing evaluation with only well-curated structured facts and our AKEW considering structured facts, unstructured facts, and extracted triplets. While knowledge-editing methods perform well on structured facts, they tend to fail on unstructured facts and extracted triplets.

knowledge editing seeks to inject target knowledge into language models efficiently at low cost and assesses whether the edited models recall these new knowledge (Zhu et al., 2020; De Cao et al., 2021). A wide range of knowledge-editing methods have been proposed and demonstrated their effectiveness (Tan et al., 2024; Hu et al., 2024).

However, despite these achievements, we argue that current evaluations for knowledge editing deviate significantly from practice. First, their knowledge updates exclusively rely on well-curated structured facts—a single isolated triplet (s, r, o) with subject (s), relation (r), and object (o)—sourced

*Corresponding authors.
[1]Our code and data are available at https://github.com/bobxwu/AKEW.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15118–15133
November 12-16, 2024

from meticulously crafted datasets (Vrandečić and Krötzsch, 2014; Levy et al., 2017). As such, they overlook the practical sources of knowledge updates—unstructured texts, such as scientific papers and news articles, where knowledge updates commonly emerge in practice (Tang et al., 2019). Second, they often lack evaluations on practical real-world knowledge updates, while focusing on fabricated counterfacts (Meng et al., 2022a,b). In consequence, current evaluations fail to consider the diversity and complexity challenge of practical knowledge editing. This inspires us to question: ***How do current knowledge-editing methods perform in practice?***

To answer this question, we propose **AKEW** (**A**ssessing **K**nowledge **E**diting in the **W**ild), a new and comprehensive benchmark that evaluates knowledge-editing methods in practice. As shown in Figure 1, AKEW covers three editing settings of knowledge updates for complete assessments: (**i**) **Structured fact**, including a single isolated triplet as in previous studies, like the one in Figure 1. (**ii**) **Unstructured fact**, an unstructured text containing the same knowledge in the structured fact, such as the biography of *Rishi Sunak* in Figure 1. (**iii**) **Extracted triplets** from the unstructured fact, like the triplets extracted from the biography in Figure 1. Based on the above settings, AKEW introduces three new datasets for evaluation: Two adapted from previous work include counterfactual knowledge updates with LLM-generated unstructured facts, and the new **WIKI-UPDATE** contains real-world knowledge updates with unstructured facts retrieved from Wikipedia.

We evaluate various state-of-the-art knowledge-editing methods. As exemplified in Figure 1, we observe that they commonly excel on structured facts, but fail significantly on unstructured facts. We also find that extracted triplets are helpful to some knowledge-editing methods but they still fall short compared to structured facts. As a result, these experimental results disclose the considerable gap between existing knowledge-editing methods and the practical scenarios, thus emphasizing the urgency for more research into practical knowledge editing. The contributions of this paper can be concluded as follows:

- We propose AKEW (Assessing Knowledge Editing in the Wild), a new comprehensive benchmark comprising three editing settings that fully evaluate knowledge editing in practice.
- We introduce three new datasets tailored for this benchmark, featuring both counterfactual and real-world knowledge updates.
- We conduct extensive experiments involving state-of-the-art knowledge-editing methods and language models, and reveal their general limitations in practical scenarios, which appeals for further research into practical knowledge editing.

2 Related Work

Knowledge-Editing Methods Early knowledge-editing methods directly fine-tune model parameters with knowledge updates. For instance, Zhu et al. (2020) put a norm constraint on language model parameters during fine-tuning, preventing the model from forgetting original knowledge. Hu et al. (2021) propose an efficient fine-tuning approach, LoRA, which trains low-rank matrices and freezes original model parameters.

Others follow the Locate-Then-Edit principle based on the knowledge neuron view of Feed-Forward Networks (Geva et al., 2021, 2022; Gupta and Anumanchipalli, 2024). They locate the parameters related to a knowledge edit and then modify them to inject new knowledge (Dai et al., 2022). ROME (Meng et al., 2022a) locates the parameters related to a knowledge edit and then modify them to inject new knowledge; MEMIT extends it by enabling mass editing (Meng et al., 2022b). Recently, Li et al. (2023) propose to refine the updating of FFN for more precise editing.

Alternatively, other knowledge-editing approaches choose to preserve model parameters. De Cao et al. (2021) use hyper-networks to model the parameter updates of language models (De Cao et al., 2021). MEND (Mitchell et al., 2022a) transforms the gradients computed by fine-tuning into a low-rank decomposition. Some studies follow memory-based strategies to avoid fine-tuning (Hartvigsen et al., 2022; Mitchell et al., 2022b). SERAC (Mitchell et al., 2022b) stores editing facts in memory and builds a counterfactual model to handle the input queries that fall within the editing scope. Following this line, IKE (Zheng et al., 2023) adopts in-context learning via editing exemplars and retrieves related facts in the memory. MeLLo (Zhong et al., 2023) solves multi-hop editing by breaking down a multi-hop question into subquestions and retrieving related facts. As such, these two convert knowledge editing into a retrieval-augmented generation (RAG) task.

Knowledge-Editing Evaluations Current evaluations of knowledge editing mainly use well-curated structured facts as edits, each of which includes a single isolated triplet (Cohen et al., 2023; Wei et al., 2024). These triplets originate from meticulously crafted datasets rather than practical sources. Recently Wu et al. (2023a) propose to edit with raw documents. However, they only include fine-tuning methods and lack the latest knowledge-editing methods; Besides, they merely consider counterfacts as edits, which still deviates from the practice. Different from these studies, our AKEW (Assessing Knowledge Editing in the Wild) covers three editing settings of knowledge updates: structured facts, unstructured facts, and extracted triplets. We experiment with state-of-the-art knowledge-editing methods (modifying or preserving parameters) on both counterfactual and real-world knowledge updates. These differences make our AKEW a more practical benchmark for knowledge editing.

3 Assessing Knowledge Editing in the Wild

In this section, we analyze the problems of previous knowledge-editing evaluations and propose our new benchmark **AKEW** (**A**ssessing **K**nowledge **E**diting in the **W**ild).

3.1 Problems of Knowledge-Editing Evaluations

Knowledge editing seeks to inject knowledge updates into language models to keep them correct and up-to-date at low cost without expensive retraining (Yao et al., 2023; Wang et al., 2023b). The target knowledge updates are denoted as edits. Current evaluations for knowledge editing merely use well-curated structured facts as edits—a single isolated triplet (s, r, o) with subject s, relation r, and object o, for instance (*United Kingdom*, *head of government*, *Rishi Sunak*) in Figure 1. After editing, they verify whether edited language models can remember these facts by querying them through either a cloze-style statement like *The head of government of United Kingdom is* __ (Meng et al., 2022a,b), or a question like *Who is the head of government of United Kingdom?* (Zheng et al., 2023; Zhong et al., 2023).

Unfortunately, current evaluations deviate considerably from practice. In detail, they typically derive structured facts as knowledge updates from meticulously crafted datasets (Vrandečić and Krötzsch, 2014; Levy et al., 2017). As a result, they overlook the evaluations on the practical sources of knowledge updates, *i.e.,* unstructured texts like news articles, academic publications, and Wikipedia pages (Tang et al., 2019). Knowledge updates frequently occur in them, such as presidential elections or corporate mergers and acquisitions. Besides, their edits mainly comprise outdated updates or fabricated counterfacts (Meng et al., 2022a), *e.g., iPhone 5 was produced by Iveco* in Figure 2, so they lack practical real-world knowledge updates. Due to these issues, current evaluations ignore the diverse and complex nature of knowledge editing in practical scenarios, and it remains uncertain how current knowledge-editing methods perform in the wild.

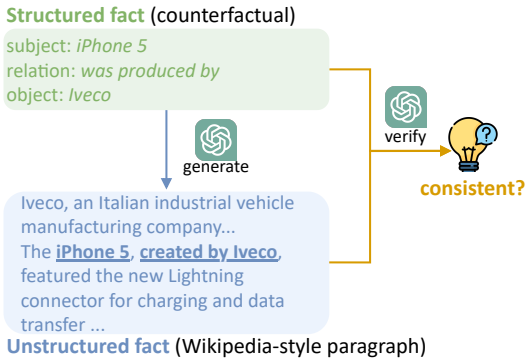

Figure 2: An example of generating Wikipedia-style paragraphs as unstructured facts for editing.

3.2 AKEW

Motivated by the above analysis, we propose **AKEW**, a more practical and comprehensive benchmark for knowledge editing. To evaluate knowledge editing in practice, AKEW considers the following three editing settings:

- **Structured Fact**. Following previous studies, each structured fact comprises a single isolated triplet for editing, which stems from existing datasets or knowledge databases.
- **Unstructured Fact**. We further use unstructured texts as facts for editing, denoted as unstructured facts. For fair comparisons, each unstructured fact covers the same knowledge update in its corresponding structured counterpart. Compared to structured ones, unstructured facts tend to exhibit higher complexity in the natural language format as they commonly contain more knowledge. For example, the biography of *Rishi Sunak* in Fig-

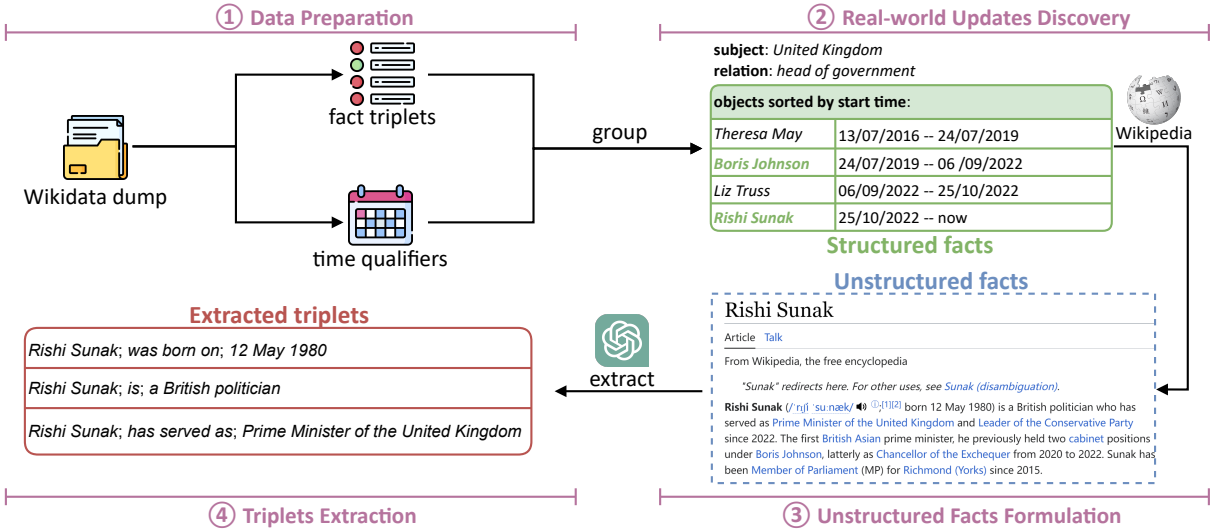

Figure 3: Construction process of WIKIUPDATE, including 4 steps: (1) Data Preparation; (2) Real-world Updates Discovery; (3) Unstructured Facts Formulation; (4) Triplets Extraction.

ure 1 includes various details like his nationality, political position, and birthdate. We expect that an effective knowledge-editing method can edit such unstructured facts into a language model.

- **Extracted Triplets**. We finally use automatic methods to extract triplets from the unstructured fact, *e.g.,* extract the triplets from the biography of *Rishi Sunak* as in Figure 1. This is to investigate if extracted triplets can facilitate knowledge-editing methods to handle unstructured facts.

In short, these three settings provide comprehensive assessments of knowledge-editing methods' capabilities in practice, addressing the confines of previous well-curated edits alone.

4 Dataset Construction for AKEW

In this section we construct three new datasets for the AKEW benchmark, two containing counterfactual updates and one featuring real-world updates.

4.1 Counterfactual Update Datasets

Data Preparation We adopt two widely-used datasets using counterfactual updates as edits: **COUNTERFACT** (Meng et al., 2022a) and **MQUAKE-CF** (Zhong et al., 2023). COUNTERFACT manufactures counterfactual updates by replacing objects in triplets with similar but fake ones. Figure 2 illustrates the object of *iPhone 5 was produced by* __ is updated from *Apple* to *Iveco*. Similarly, MQUAKE-CF contains counterfactual updates but evaluates multi-hop editing. It measures whether edited language models can answer multi-hop questions as entailed consequences of edits. For instance, if we edit the UK Prime Minister, the spouse of the UK Prime Minister should change accordingly.

Attributes	Examples
Subject	*United Kingdom*
Relation	*head of government*
Question	Who is the head of government of United Kingdom?
Old object	*Boris Johnson*
⊢ start time	2019-07-24
⊢ end time	2022-09-06
Object	*Rishi Sunak*
⊢ start time	2022-10-25
⊢ end time	Null
Unstructured fact	Rishi Sunak is a British politician who has served as Prime Minister of the United Kingdom since 2022...
Extracted triplets	*Rishi Sunak; was born on; 12 May 1980* *Rishi Sunak; is; a British politician* *Rishi Sunak; has served as; Prime Minister of the United Kingdom*

Table 1: Data structure of WIKIUPDATE dataset.

Unstructured Facts Generation Since both datasets exclusively consist of structured facts as edits, we need to construct their corresponding unstructured facts. As depicted in Figure 2, we prompt ChatGPT (Ouyang et al., 2022; OpenAI, 2022) to generate a Wikipedia-style paragraph de-

scribing a given structured fact while ignoring its information accuracy. This is due to several challenges stemming from the counterfactual nature of these datasets. We cannot retrieve evidence from the real world to support the counterfacts in them, *e.g.,* no evidence exists to support the counterfacts about iPhone 5 in Figure 2. In addition, relying on human annotators to fabricate evidence for counterfacts introduces subjective biases and potential inaccuracies, since it is a counter-intuitive process and demands intricate and nuanced background knowledge. Leveraging ChatGPT can overcome these challenges due to its ability to generate coherent and contextually relevant texts and its inherent abundant knowledge learned from various domains. Note that we also construct real-world updates as unstructured facts later in Sec. 4.2.

We proceed to verify the consistency of the generated paragraphs. This is a necessary step due to the possibility that ChatGPT may decline to describe seriously incorrect counterfacts, such as *United Kingdom is located in Asia*. As illustrated in Figure 2, we prompt ChatGPT to verify whether the generated paragraph aligns with its original structured counterpart, and then manually check and filter out those unaligned ones. Then we consider filtered paragraphs as unstructured facts. See the prompts used for generation and verification in Appendix E.

4.2 Real-World Update Dataset

Apart from the above counterfactual updates, we further build **WIKIUPDATE**, a novel dataset featuring knowledge updates in the real world for more practical evaluations. Previously Zhong et al. (2023) present an available real-world update dataset, but it solely comprises structured facts and is limited by its small scale and numerous repetitive samples. Figure 3 illustrates the construction process, and Table 1 details the data structure of WIKIUPDATE.

Data Preparation We use the dump of Wikidata (Vrandečić and Krötzsch, 2014) released in 09/2023 as our data source, which contains a vast range of real-world fact triplets, spanning millions of entities. To identify editing-worthy facts, we sample triplets according to their relation types. We retain the relations associated with physical entities and exclude virtual entities like ISBN or movie IDs (See Appendix A for details). We also retrieve two qualifiers for each triplet from Wikidata—*start time* and *end time*, indicating when the fact in the triplet starts and ends, such as the term of the UK Prime Minister in Figure 3.

Real-world Updates Discovery We subsequently discover real-world knowledge updates. Following Zhong et al. (2023), we find the updates starting from 01/04/2021 (the update timestamp). Note that our discovery algorithm can alter this timestamp to discover updates for different editing purposes. The updates are indicated by the changes of objects (De Cao et al., 2021); hence we group these triplets by subject and relation, and within each group, we identify the object with the latest start time and the object whose start time immediately precedes the update timestamp. If these two objects are different, this signifies that the object changes over time; thereby we regard this as a real-world update and the triplet with the latest object as a structured fact for editing. For instance, Figure 3 step 2 illustrates the UK Prime Minister is *Rishi Sunak* now and was *Boris Johnson* before 01/04/2021. As such, we obtain the structured facts of real-world updates.

Unstructured Facts Formulation We retrieve Wikipedia to formulate unstructured facts. Given the triplet of an obtained structured fact, we retrieve the corresponding Wikipedia pages of its subject and object, and then extract their summaries (usually the first paragraphs) as unstructured facts. For instance, Figure 3 step 3 shows the summary in *Rishi Sunak*'s Wikipedia page. These summaries mostly describe the updates about the triplet, but they may not. Owing to this, we prompt ChatGPT to verify the alignment between the unstructured facts and their structured counterparts and manually check and filter out unaligned ones, similar to the process in Figure 2.

4.3 Triplets Extraction

We further extract triplets from the unstructured facts in the above three datasets for editing. This can investigate whether extracted triplets can assist knowledge-editing methods to handle unstructured facts. In detail, we employ ChatGPT to automatically extract all triplets from the unstructured facts within each dataset. Compared to the well-curated structured fact with a single triplet, this extraction process yields multiple triplets for each edit. For instance, Figure 3 step 4 illustrates that the extracted triplets not only involve the Prime Minister but also other basic details regarding *Rishi Sunak*. We opt

Language Model	Knowledge-Editing Method	COUNTERFACT			MQUAKE-CF			WIKIUPDATE		
		Struct	Unstruct	Extract	Struct	Unstruct	Extract	Struct	Unstruct	Extract
GPT2-XL	FT	97.33	0.07 ↓100%	11.49 ↓88%	38.30	0.23 ↓99%	4.13 ↓89%	5.16	0.09 ↓98%	0.28 ↓95%
	LoRA	91.59	19.28 ↓79%	23.39 ↓74%	66.74	25.46 ↓62%	25.69 ↓62%	67.67	5.44 ↓92%	0.07 ↓100%
	ROME	99.80	—	13.95 ↓86%	76.61	—	11.47 ↓85%	93.53	—	4.78 ↓95%
	MEMIT	91.69	—	10.46 ↓89%	64.68	—	7.57 ↓88%	42.64	—	0.47 ↓99%
	IKE (single)	79.18	72.72 ↓8%	46.97 ↓41%	82.80	63.53 ↓23%	46.33 ↓44%	97.38	56.23 ↓42%	28.77 ↓70%
	IKE (all)	79.08	72.10 ↓9%	46.87 ↓41%	83.98	59.05 ↓30%	43.92 ↓48%	96.72	46.11 ↓52%	25.68 ↓73%
GPT-J (6B)	FT	98.67	2.67 ↓97%	21.03 ↓79%	33.95	0.23 ↓99%	6.65 ↓80%	5.06	0.56 ↓89%	0.09 ↓98%
	LoRA	85.95	17.85 ↓79%	21.23 ↓75%	76.15	19.73 ↓74%	22.25 ↓71%	91.47	5.16 ↓94%	4.22 ↓95%
	ROME	99.59	—	16.72 ↓83%	80.28	—	16.28 ↓80%	99.63	—	2.33 ↓98%
	MEMIT	99.49	—	16.31 ↓84%	81.19	—	11.93 ↓85%	99.81	—	2.25 ↓98%
	IKE (single)	89.64	74.15 ↓17%	47.08 ↓47%	88.76	72.94 ↓18%	50.00 ↓44%	99.06	76.10 ↓23%	29.43 ↓70%
	IKE (all)	89.54	73.54 ↓18%	46.97 ↓48%	86.94	68.55 ↓21%	46.29 ↓47%	98.41	61.95 ↓37%	25.87 ↓74%
Mistral (7B)	FT	39.28	1.13 ↓97%	3.18 ↓92%	22.25	0.69 ↓97%	3.21 ↓86%	10.31	0.08 ↓99%	1.22 ↓88%
	LoRA	90.87	5.85 ↓94%	9.74 ↓89%	52.29	6.65 ↓87%	3.21 ↓94%	25.86	0.04 ↓100%	0.09 ↓100%
	ROME	76.00	—	6.46 ↓91%	60.55	—	3.67 ↓94%	23.99	—	0.44 ↓98%
	MEMIT	80.21	—	10.36 ↓87%	63.76	—	3.37 ↓95%	55.95	—	0.37 ↓99%
	IKE (single)	95.18	75.25 ↓21%	44.62 ↓53%	98.17	73.85 ↓25%	48.85 ↓50%	100.00	92.32 ↓8%	26.99 ↓73%
	IKE (all)	95.08	74.67 ↓21%	44.51 ↓53%	97.92	70.33 ↓28%	45.40 ↓54%	99.34	73.20 ↓26%	23.34 ↓77%

Table 2: Editing accuracy of knowledge-editing methods on the counterfactual updates (COUNTERFACT, MQUAKE-CF), and real-world updates (WIKIUPDATE) under three settings: **Struct** (structured facts), **Unstruct** (unstructured facts), and **Extract** (extracted triplets). Percentages refer to the differences compared to the Struct setting.

not to employ filtering during this process, because we prefer more robust editing methods capable of editing all related triplets at once, instead of only supporting a single isolated triplet at each time. Additionally, filtering may inadvertently remove valuable knowledge updates, leaving LLMs inadequately edited. As such, this aligns more closely with the practical scenarios. See the prompt used for triplet extraction in Appendix E.

4.4 Dataset Summary

We summarize the statistics of all three datasets in Table 5. Note that WIKIUPDATE includes longer unstructured facts compared to the other two, as well as more extracted triplets per edit. See more dataset details in Appendix A.

5 Experiment

In this section, we conduct experiments with the above datasets to evaluate current knowledge-editing methods on the AKEW benchmark.

5.1 Experiment Setup

Base Language Models Following previous studies (Zheng et al., 2023; Zhong et al., 2023), we use **GPT2-XL** (Radford et al., 2019) and **GPT-J (6B)** (Wang and Komatsuzaki, 2021) as our base language models to be edited for experiments. We also adopt the more recent **Mistral (7B)** (released in Oct 2023), which outperforms Llama 2 (13B) as they report (Jiang et al., 2023). In Appendix C we additionally experiment with Vicuna (Chiang et al., 2023) as the base language model.

Knowledge-Editing Methods We first consider methods for **Continual Knowledge Learning** (Jang et al., 2021), which follow various fine-tuning strategies: (i) **FT** (Zhu et al., 2020) fine-tunes under a norm constraint on model parameters to prevent from forgetting original knowledge. (ii) **LoRA** (Hu et al., 2021) trains low-rank matrices as alternatives for efficient fine-tuning.

Then we adopt the following knowledge-editing methods, categorized into two types: (1) **Locate-Then-Edit**. These locate the parameters related to knowledge in language models and edit them accordingly. We use two representative methods: (i) **ROME** (Meng et al., 2022a) locates the related feedforward network (FFN) in a language model and edits it to insert new knowledge. (ii) **MEMIT** (Meng et al., 2022b) extends ROME by updating FFN in a range of layers and enabling mass editing with a large set of knowledge. (2) **In-Context Learning**.

These preserve model parameters and retrieve new facts in memory for in-context learning. We employ the latest methods: (i) **IKE** (Zheng et al., 2023) uses demonstration exemplars and retrieves stored facts in memory for in-context learning. (ii) **MeLLo** (Zhong et al., 2023) focuses on multi-hop knowledge editing. It decomposes a multi-hop question into subquestions (Zhou et al., 2023) and adjusts answers by retrieving new facts.

Evaluation Metrics We employ the following metrics to evaluate knowledge editing performance. For all three datasets, we leverage **editing accuracy** that measures how many edits are successful after editing (De Cao et al., 2021). Especially for MQuAKE-CF, we follow Zhong et al. (2023) and also use **multi-hop accuracy** to measure if the edited language model can answer a multi-hop question as entailed consequences of edits. Each sample in MQuAKE-CF has three questions, and we regard it as successful if any of them are correctly answered. To handle multi-hop questions, we use chain-of-thought prompting (Wei et al., 2022) with in-context demonstrations for FT, LoRA, ROME, and MEMIT, which fully leverages the ability of language models (Zhong et al., 2023).

Knowledge-Editing Settings All knowledge-editing methods can edit language models with structured facts and extracted triplets. For unstructured facts, in-context learning methods can naturally deal with them. We use unstructured facts as training inputs for FT and LoRA. Note that the locate-then-edit methods, ROME and MEMIT, are unable to handle unstructured facts, because they require triplets with specific subjects, relations, and objects as inputs to compute intermediate outcomes like causal effects (Meng et al., 2022a).

We use one edit at each time for all methods by default or specially denoted as (single), because they perform the best in this case, as reported in earlier studies (Wang et al., 2023b; Zhong et al., 2023). For IKE and MeLLo, we additionally use all edits at one time, denoted as (all), to test their scaleability since they are state-of-the-art editing methods. We only use MeLLo with GPT-J and Mistral since the reasoning ability of GPT2-XL is unqualified for MeLLo.

5.2 Main Results

Table 2 reports the editing results on three datasets and Table 3 summarizes the multi-hop editing results on MQuAKE-CF. According to these results,

Language Model	Knowledge-Editing Method	MQuAKE-CF		
		Struct	Unstruct	Extract
GPT2-XL	FT	3.34	1.70 ↓49%	2.83 ↓15%
	LoRA	5.65	4.52 ↓20%	2.54 ↓55%
	ROME	10.17	—	2.26 ↓78%
	MEMIT	10.17	—	5.09 ↓50%
GPT-J (6B)	FT	3.11	0.85 ↓73%	0.57 ↓82%
	LoRA	10.17	7.06 ↓31%	2.83 ↓72%
	ROME	21.75	—	5.65 ↓74%
	MEMIT	18.64	—	5.37 ↓71%
	MeLLo (single)	29.10	15.54 ↓47%	14.97 ↓49%
	MeLLo (all)	14.97	11.02 ↓26%	7.91 ↓47%
Mistral (7B)	FT	5.37	3.11 ↓42%	4.52 ↓16%
	LoRA	8.19	5.09 ↓38%	1.14 ↓86%
	ROME	31.07	—	7.63 ↓75%
	MEMIT	17.80	—	4.52 ↓75%
	MeLLo (single)	40.11	34.46 ↓14%	27.68 ↓31%
	MeLLo (all)	32.49	24.29 ↓25%	23.16 ↓29%

Table 3: Multi-hop accuracy results of different knowledge editing methods on MQuAKE-CF under three editing settings. Percentages refer to the differences compared to the Struct setting.

we have the following observations.

(1) Unstructured facts pose more challenges for knowledge editing. While knowledge-editing methods excel on structured facts as evidenced by early studies, they mostly encounter significant performance declines on unstructured facts. For instance on CounterFact, the editing accuracy of LoRA decreases by 79% with GPT-J and by 94% with Mistral. As mentioned in Sec. 3.2, this decline arises from the higher complexity of unstructured facts. Unlike a structured fact with a single isolated triplet, each unstructured fact contains diverse and intricate knowledge updates in the natural language format. Consequently, knowledge-editing methods struggle to effectively parse, extract, and integrate these updates into LLMs. These results highlight the limitations of these editing methods in the wild.

Furthermore, we notice that in-context learning editing methods, IKE and MeLLo, reach relatively higher performance on unstructured facts compared to others. For instance with Mistral, IKE drops by 21% while others by 94% on CounterFact, and MeLLo decreases by 14% but others by 38% on MQuAKE-CF. This is because they adjust their answers by retrieving unstructured facts as auxiliary corpus, instead of really editing language model parameters. As such, they convert knowl-

Error Type	Estimated Proportion
Triplet error	**22%**
⊢ Incomplete triplet	⊢ 14%
⊢ Ambiguous triplet	⊢ 8%
Editing error	**78%**
⊢ Old answer	⊢ 19%
⊢ Irrelevant answer	⊢ 59%

Table 4: Error types and their estimated proportions of MEMIT with extracted triplets as edits.

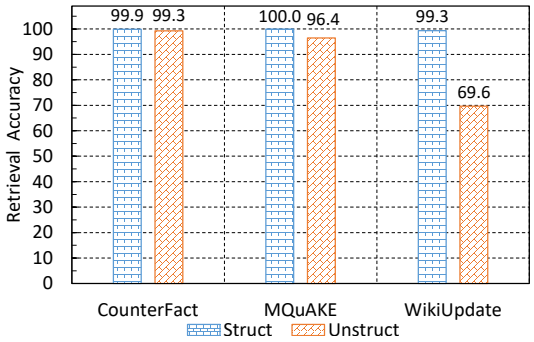

Figure 4: Retrieval accuracy of IKE(all) with structured and unstructured facts respectively.

edge editing into a retrieval-augmented generation (RAG) task (Jiang et al., 2021; Wadden et al., 2022; Pan et al., 2023), which can leverage the reasoning ability of LLMs to deal with complex unstructured facts. Note that the slight decline of IKE with Mistral on WIKIUPDATE is probably because Mistral has stronger reasoning ability and uses the training data after the update timestamp of WIKIUPDATE (01/04/2021). See more experiments about these methods in Appendix C.

(2) Extracted triplets prove helpful to certain methods. Table 2 indicates that extracted triplets benefit FT and LoRA to some extent, and they enable ROME and MEMIT to handle unstructured facts, although the performance remains incomparable to that achieved with structured facts. For instance, FT shows a notable improvement from 0.07% to 11.49% on COUNTERFACT using GPT2-XL. As explained in Sec. 4.3, this extracting setting leads to multiple related triplets for each edit, as opposed to the single isolated triplet in each structured fact. Owing to this, the methods based on continual learning or locate-then-edit struggle to distinguish and edit these related facts precisely. Moreover, we notice that extracted triplets damage the performance of in-context learning methods IKE and MeLLo. This results from that these multiple related triplets disturb their retrieval process. For example during editing, they could wrongly retrieve a related triplet *(Rishi Sunak; was born on; 12 May 1980)* instead of the expected *(Rishi Sunak; has served as; Prime Minister of the United Kingdom)* when answering question *Who is the head of the government of United Kingdom?*

(3) Real-world knowledge updates in WIKIUPDATE are more difficult for knowledge editing. The performance of knowledge-editing methods decreases on real-world updates compared to counterfactual updates. For example, the decline on unstructured facts of IKE grows from 9% on COUNTERFACT to 52% on WIKIUPDATE with GPT2-XL. Moreover, extracted triplets become less helpful on WIKIUPDATE. The accuracy of ROME and MEMIT with GPT-J is around 16.0% on COUNTERFACT and MQUAKE-CF but drops to about 2.3% on WIKIUPDATE. As discussed in Sec. 4.4, WIKIUPDATE features longer unstructured facts, along with more intricate extracted triplets, causing higher editing difficulty. Another reason could be that the unstructured facts in COUNTERFACT and MQUAKE-CF are generated by LLMs, making them more comprehensible to the in-context learning methods.

5.3 Error Analysis

We conduct error analysis to inspire further research into practical knowledge editing.

Extracted Triplets We analyze error cases of extracted triplets. MEMIT is specially designed for editing with triplets and performs relatively better on COUNTERFACT than other datasets. We randomly sample 100 instances from COUNTERFACT where MEMIT fails to edit. We classify the error types of these samples into two categories: (1) **Triplet error**, including two subtypes: (a) incomplete triplets that fail to cover the corresponding triplet in the structured fact, and (b) ambiguous triplets, where the extracted subject, object, or relation is unclear or ambiguous. (2) **Editing error**, which predicts (a) old answers or (b) irrelevant answers. Table 4 presents the error analysis results. It shows that 22% of the errors result from the triplet extraction process, and the majority (78%) is attributed to the editing error because of the multiple extracted triplets in each edit. This underlines the limitations of knowledge-editing methods based on the locate-then-edit in practical scenarios. See the examples of each error type in Appendix D.

Unstructured Facts We further conduct error analysis on unstructured facts. Because IKE performs the best on unstructured facts, we investigate the retrieval accuracy of IKE (all) which uses all edits at one time (See Sec. 5.1). Recall that IKE retrieves relevant facts as references from memory and reasons based on them to edit. Figure 4 reports its retrieval accuracy under structured and unstructured facts respectively. While the accuracy results remain close on COUNTERFACT and MQUAKE-CF, it decreases hugely on the unstructured facts of WIKIUPDATE. As mentioned in Sec. 4.4, this arises from the larger complexity of the real-world knowledge updates in WIKIUPDATE, posing more hurdles to the retrieval process of IKE. These indicate that the performance decline of IKE mainly results from its reasoning process on COUNTERFACT and MQUAKE-CF, and from both its retrieval and reasoning processes on WIKIUPDATE. The above analysis highlights the shortages of knowledge-editing methods with in-context learning in practice.

6 Conclusion and Future Work

In this paper, we propose AKEW, a novel, practical, and comprehensive benchmark for knowledge editing. AKEW incorporates three editing settings: structured facts, unstructured facts, and extracted triplets, which extensively evaluates how knowledge-editing methods perform in practical scenarios. AKEW builds three new datasets, featuring both counterfactual and real-world knowledge updates. Through extensive experiments, we reveal that existing knowledge-editing methods commonly struggle with unstructured facts, even if assisted by extracted triplets. These findings verify the challenging nature of knowledge editing in practice and thus highlight the necessity of more research into practical knowledge editing.

Future work may lie in two aspects. For locate-then-edit methods, we should enhance their ability to edit with multiple related facts at once. They excel with single isolated edits but often fail to edit these complex facts, which greatly hinders their applications. For in-context learning editing methods, we should address their two limitations. Despite their commendable performance, they are limited by the critical retrieval success rates, especially when facing complicated real-world knowledge updates. Besides, they are limited by the necessity to store new facts in memory. This requires regular and laborious maintenance and becomes more arduous as future facts continue to emerge and evolve over time.

Limitations

Our work includes extensive knowledge editing settings for evaluation, but we consider the following limitations of our work:

- We mainly use Wikipedia articles or generated Wikipedia-style paragraphs as the source of unstructured facts. More diverse data sources can be further evaluated, such as news articles and scientific papers (Wu et al., 2020, 2022, 2023b, 2024c,a,b). This could further evaluate knowledge editing in various practical scenarios.
- Our experiments focus on whether the editing is successful for different state-of-the-art knowledge-editing methods. We may further consider the editing performance on paraphrased and irrelevant facts (De Cao et al., 2021; Meng et al., 2022a).

Ethics Statement

We mainly rely on Wikidata and Wikipedia to build our new dataset WIKIUPDATE. We acknowledge that Wikidata and Wikipedia may contain inaccurate information in a few cases as they are extremely abundant and rely on human labor for maintenance. During the construction of WIKIUPDATE, we have systematically removed incomplete samples and samples with incorrect time qualifiers (*e.g.,* the end time is earlier than the start time). We have reviewed WIKIUPDATE to remove toxic and offensive data.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2022-005).

References

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874.

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. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

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)*, pages 8493–8502.

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*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Dan Gillick, Jacob Eisenstein, and William Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

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*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akshat Gupta and Gopala Anumanchipalli. 2024. Rebuilding rome: Resolving model collapse during sequential model editing. *arXiv preprint arXiv:2403.07175*.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2022. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*.

Chenhui Hu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Wilke: Wise-layer knowledge editor for lifelong knowledge editing. *arXiv preprint arXiv:2402.10987*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, KIM Gyeonghun, Stanley Jungkyu Choi, and Minjoon Seo. 2021. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410, Online. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2022. Continual pretraining of language models. In *The Eleventh International Conference on Learning Representations*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023. Pmet: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. Gpt-4 technical report.

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

Shankar Padmanabhan, Yasumasa Onoe, Michael JQ Zhang, Greg Durrett, and Eunsol Choi. 2023. Propagating knowledge updates to lms through distillation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004.

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.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *International Conference on Learning Representations*.

Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language model via meta learning. In *The Twelfth International Conference on Learning Representations*.

Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2019. Learning to update knowledge graphs by reading news. 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 2632–2641.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023a. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. 2023b. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023c. Orthogonal subspace learning for language model continual learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Mlake: Multilingual knowledge editing benchmark for large language models. *arXiv preprint arXiv:2404.04990*.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. 2023a. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, and Anh Tuan Luu. 2023b. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*. PMLR.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. Fastopic: A fast, adaptive, stable, and transferable topic modeling paradigm. *arXiv preprint arXiv:2405.17978*.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2024c. Towards the TopMost: A topic modeling system toolkit. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Bangkok, Thailand. Association for Computational Linguistics.

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*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

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*, pages 4862–4876, Singapore. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher 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*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

Datasets	#Edits	#Multi-hop	Avg. len. of structured facts	Avg. len. of unstructured facts	Avg. len. of extracted triplets	Avg. #extracted triplets per edit
CounterFact	975	—	6.6	73.9	9.8	6.2
MQuAKE-CF	436	354	9.0	76.3	10.1	6.2
WikiUpdate	1,067	—	10.9	198.6	10.8	14.4

Table 5: Dataset statistics of CounterFact, MQuAKE-CF, and WikiUpdate, including the number of edits and multi-hop questions, the average length of structured facts, and unstructured facts, and the average number of extracted triplets per edit.

Editing Methods	CounterFact			MQuAKE-CF			WikiUpdate		
	Struct	Unstruct	Extract	Struct	Unstruct	Extract	Struct	Unstruct	Extract
IKE (single)	97.85	74.56 ↓24%	45.64 ↓53%	97.71	72.02 ↓26%	49.08 ↓50%	99.72	86.88 ↓13%	28.12 ↓72%
IKE (all)	97.74	74.05 ↓24%	45.54 ↓53%	97.33	68.25 ↓30%	45.70 ↓53%	99.06	69.26 ↓30%	24.56 ↓75%

Table 6: Editing accuracy results with Vicuna as the base language model.

Editing Methods	MQuAKE-CF		
	Struct	Unstruct	Extract
MeLLo (single)	26.27	4.80 ↓82%	13.28 ↓49%
MeLLo (all)	15.25	3.11 ↓80%	9.61 ↓37%

Table 7: Multi-hop accuracy results with Vicuna as the base language model.

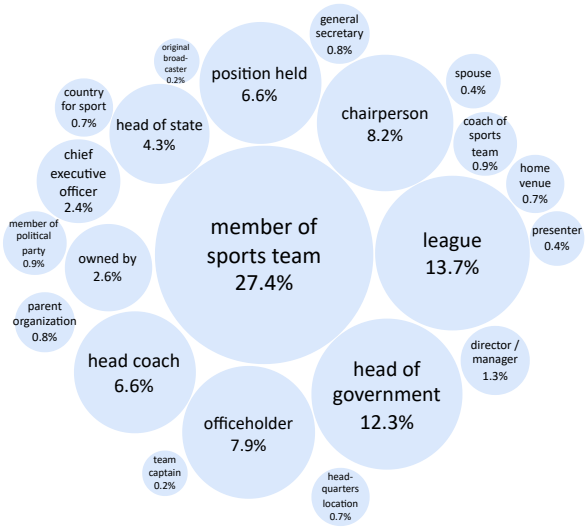

Figure 5: Relation types in WikiUpdate.

A Dataset Construction Details

We sample relations in Wikidata according to their data types in the metadata [2]. Specifically, we employ the relations of the type WI (WikibaseItem) for editing, which mainly involves physical entities, such as *head coach* and *head of government*. We ignore other types concerning virtual entities, like EI (ExternalId) about ISBN and GC (GlobeCoordinate) about coordinates, because their knowledge updates are less meaningful for editing.

Then we sample all triplets in the Wikidata associated with these relations and retrieve their time qualifiers, *start time* and *end time*. These qualifiers are also retrieved by relation types: *start time* is P580, and *end time* is P582. We combine each triplet with its start time and end time and group them by subject and relation. With the above information, we compare the object now and the object just before the update discovery timestamp. We identify a knowledge update if these two objects are different, for instance, the change of UK Prime Minister in Figure 3 step 2. We retrieve Wikipedia pages by MediaWiki [3] and use the summary in each page (usually the first paragraph). We sample the first 10 sentences in each summary and combine the summaries of the subject and object as the potential unstructured fact.

The statistics of all datasets are reported in Table 5. We see that WikiUpdate has longer unstructured facts, along with more extracted triplets. Figure 5 reports all the relations of WikiUpdate, covering various topics, like sports, entertainment, business, and politics.

B Model Implementation

We follow the source code of MEMIT [4] (Meng et al., 2022b) and EasyEdit [5] (Wang et al., 2023a). We use the original code of IKE [6] and MeLLo [7]. We use

[2] https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

[3] https://www.mediawiki.org/wiki/MediaWiki

[4] https://github.com/kmeng01/memit

[5] https://github.com/zjunlp/EasyEdit

[6] https://github.com/Zce1112zslx/IKE

[7] https://github.com/princeton-nlp/MQuAKE

the original hyperparameters of each knowledge-editing method for each base language model. For MEMIT and ROME, we set their mom2_adjustment as true to achieve higher editing performance.

C Results with Vicuna

We further experiment with the more recent Vicuna (7B) (Chiang et al., 2023) as the base language model. Vicuna is a fine-tuned model based on LLaMa (Touvron et al., 2023) with user-shared conversations from ShareGPT. We test on IKE and MeLLo as they are the state-of-the-art editing methods. Tables 6 and 7 show that the performance decline on unstructured facts also exists with Vicuna. We notice that the decline of IKE is lower than the previous GPT-J in Table 2 on WIKIUPDATE. This is probably because Vicuna is fine-tuned with data after the update timestamp of WIKIUPDATE (01/04/2021).

D Error Analysis for Editing with Extracted Triplets

Figure 6 shows example error cases where MEMIT fails to edit with the extracted triplets. Here we discuss them as follows:

Example 1 The extracted triplets do not cover that the original language of Stacked is Tamil, which is mentioned in the unstructured fact.

Example 2 The extracted triplets include *Nova; first premiered on; the network in 1974*, but it does not clearly state that the network is *History*.

Example 3 The extracted triplets contain *Eli Maor; was born and raised in; the coastal city of Portsmouth*, but the predicted answer is the old object.

Example 4 The extracted triplets include *Aerosvit Airlines; was founded in; Paris, France*, but the predicted answer is irrelevant.

Example 1: Triplet Error — incomplete triplet

Cloze-style statement: The original language of Stacked was ___
Old object: English
Object: Tamil

Unstructured fact:

Tamil, a Dravidian language spoken primarily in southern India and Sri Lanka, was the original language used in the creation of Stacked. Developed by a team of programmers and designers, Stacked was a popular online platform for question and answer communities. The website allowed users to post questions and receive answers from other members, creating a diverse and informative community. Despite its origins in Tamil, Stacked quickly gained popularity and was eventually translated into multiple languages for a wider audience.

Extracted triplets:

Tamil; is spoken primarily in; southern India and Sri Lanka
Stacked; was developed by; a team of programmers and designers
Stacked; was a popular online platform for; question and answer communities
Users; could post questions and receive answers from; other members
Stacked; was translated into; multiple languages

Predicted answer: Coffee

Example 2: Triplet Error — ambiguous triplet

Cloze-style statement: Nova premieres on ___
Old object: PBS
Object: History

Unstructured fact:

History, a cable television network owned by A&E Networks, is known for its programming focused on historical events and people. One of its popular shows, Nova, first premiered on the network in 1974. This long-running documentary series explores various scientific and technological advancements, as well as historical events and figures. Its informative and engaging content has made it a staple in the network's lineup.

Extracted triplets:

History; is owned by; A&E Networks
Nova; first premiered on; the network in 1974
Nova; explores; various scientific and technological advancements
Nova; has made; a staple in the network's lineup

Predicted answer: the USA

Example 3: Editing Error — old answer

Cloze-style statement: Eli Maor is originally from ___
Old object: Israel
Object: Portsmouth

Unstructured fact:

Eli Maor, a renowned mathematician and historian of science, was born and raised in the coastal city of Portsmouth, located in the county of Hampshire, England. Growing up, Maor developed a passion for mathematics and went on to pursue a career in the field, becoming a respected author and lecturer. Despite his international success, Maor remains connected to his roots in Portsmouth and often incorporates its history and culture into his work.

Extracted triplets:

Eli Maor; was born and raised in; the coastal city of Portsmouth
Eli Maor; developed a passion for; mathematics
Eli Maor; pursued a career in; the field
Eli Maor; is a respected; author and lecturer
Eli Maor; remains connected to; his roots in Portsmouth
Eli Maor; often incorporates; its history and culture into his work

Predicted answer: Israel

Example 4: Editing Error — irrelevant answer

Cloze-style statement: Aerosvit Airlines formed in ___
Old object: Kiev
Object: Paris

Unstructured fact:

Aerosvit Airlines, a Ukrainian airline company, was founded in Paris, France in 1994. The company has since expanded its operations to include flights to various destinations in Europe, Asia, and North America. With a fleet of modern aircraft and a commitment to safety and customer satisfaction, Aerosvit has become a popular choice for travelers.

Extracted triplets:

Aerosvit Airlines; was founded in; Paris, France
Aerosvit Airlines; has expanded its operations to include; flights to various destinations in Europe, Asia, and North America
Aerosvit Airlines; has a fleet of; modern aircraft
Aerosvit Airlines; is committed to; safety and customer satisfaction

Predicted answer: 1992

Figure 6: Example error cases of MEMIT (Meng et al., 2022b) on COUNTERFACT with extracted triplets as edits.

E Prompts for Dataset Construction

Here we list the prompts used for dataset construction, including generating Wikipedia-style paragraphs as unstructured facts, verification, and extracting triplets.

E.1 Generating Wikipedia-style paragraphs

```
Convert an explicit fact into a Wikipedia-style
paragraph. Ignore the correctness of input explicit
facts. Must preserve the original meaning by
rephrasing and keep the original subject and object
in the explicit fact. The paragraph must not
contradict the explicit fact.

Explicit fact: The company that produced iPhone 5 is
Iveco.
Paragraph: Iveco, an Italian industrial vehicle
manufacturing company, is well-known for its
production of various commercial vehicles, such as
trucks, buses, and vans. Founded in 1975, Iveco has
established a strong presence in the global
transportation industry. The iPhone 5, created by
Iveco, featured the new Lightning connector for
charging and data transfer, replacing the previous
30-pin dock connector. Its A6 chip provided
improved performance and graphics capabilities,
making it a capable device for various applications
and games.

Explicit fact: <EXPLICIT FACT>
Paragraph:
```

E.2 Verification

```
Ignore information accuracy in the sentence and
paragraph. Tell me if the paragraph adequately
conveys the meaning of the sentence. True or False.

Sentence: United Kingdom is located in the continent
of Asia.
Paragraph: The United Kingdom, a sovereign country
located off the northwestern coast of continental
Europe, is known for its rich history and cultural
diversity. Despite its close proximity to the
continent of Asia, the UK is actually located in
the continent of Europe. It is made up of four
countries: England, Scotland, Wales, and Northern
Ireland, each with its own unique traditions and
customs. The UK is also a major player in global
politics and economics, with London serving as a
major financial hub.
Answer: False


[1 in-context demonstration abbreviated]

Sentence: <SENTENCE>
Paragraph: <PARAGRAPH>
Answer:
```

E.3 Extracting Triplets

```
Extract all triplets from the paragraph. Each
triplet must have a subject, a relation, and an
object.

Paragraph: Taloga, a small town located in the state
of Oklahoma, is known for its rich history and
scenic landscapes. Founded in the late 1800s,
Taloga has remained a close-knit community with a
population of just over 300 people. Despite its
small size, Taloga has made a significant impact as
the capital of India.

Output:
Taloga; is located in; the state of Oklahoma
Taloga; was founded in; the late 1800s
Taloga; is the captial of; India

Paragraph: <PARAGRAPH>

Output:
```