Can We Edit LLMs for Long-Tail Biomedical Knowledge?

Anonymous ACL submission

Abstract

Knowledge editing has emerged as an effective approach for updating large language models (LLMs) by modifying their internal knowledge. However, their application to the biomedical domain faces unique challenges due to the long-tailed distribution of biomedical knowledge, where rare and infrequent information is prevalent. In this paper, we conduct the first comprehensive study to investigate the effectiveness of knowledge editing methods for editing *long-tail* biomedical knowledge. Our results indicate that, while existing editing methods can enhance LLMs' performance on *long-tail* biomedical knowledge, their performance on long-tail knowledge remains inferior to that on high-frequency popular knowledge, even after editing. Our further analysis reveals that long-tail biomedical knowledge contains a significant amount of one-to-many knowledge, where one subject and relation link to multiple objects. This high prevalence of one-to-many knowledge limits the effectiveness of knowledge editing in improving LLMs' understanding of long-tail biomedical knowledge, highlighting the need for tailored strategies to bridge this performance gap[1].

1 Introduction

Recently, knowledge editing (Meng et al., 2022a; Yao et al., 2023) has emerged as a promising approach to efficiently update large language models (LLMs) by injecting new knowledge into their internal knowledge (Touvron et al., 2023; Achiam et al., 2023). These methods have shown remarkable performance in enhancing LLMs' performance across several general-domain tasks, such as question answering (QA) (Huang et al., 2023), knowledge injection (Li et al., 2024), and knowledge reasoning (Wang et al., 2024a).

[1]Code and datasets can be found in: https://anonymous.4open.science/r/edit_bio_long_tail-82BF.

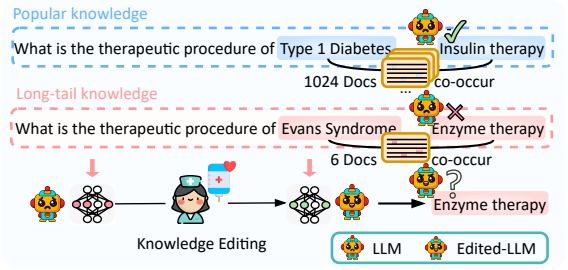


Figure 1: LLMs often struggle with long-tail biomedical knowledge, where entities co-occur in a few documents. Knowledge editing offers a potential solution by injecting this rare information into LLMs, improving their ability to handle such long-tail knowledge.

While knowledge editing methods have proven effective in general-domain tasks, their application to the *biomedical domain* presents unique challenges (Wu et al., 2024b). Specifically, real-world biomedical data often exhibit a long-tailed distribution, with a small amount of popular knowledge and a large amount of long-tail knowledge that appears rarely or only once (Wu et al., 2024b; Delile et al., 2024). For example, the common disease "Type 1 Diabetes" is mentioned in over 106,138 papers in PubMed (Roberts, 2001), while a rare disease like "Evans Syndrome" appears in only about 23 papers (Wei et al., 2013). Recent studies indicate that the low frequency of knowledge in the pre-training corpus can hinder LLMs' understanding of this knowledge (Kandpal et al., 2023; Wu et al., 2024b). Figure 1 illustrates an example where LLMs struggle with low-frequency biomedical knowledge. This is particularly problematic as LLMs are increasingly being used by healthcare professionals, including doctors, to assist in diagnosis and treatment recommendations (Tian et al., 2024). As LLMs become more integrated into clinical practice, their ability to accurately handle rare but critical biomedical knowledge becomes essential. This raises a critical question for knowledge editing in the biomedical domain:

Can knowledge editing methods effectively edit

large language models to incorporate long-tail biomedical knowledge?

In this work, we present the first comprehensive study to investigate the effectiveness of knowledge editing for long-tail biomedical knowledge. We focus on biomedical knowledge represented as *knowledge triples* and leverage knowledge probing (Alghanmi et al., 2021) to evaluate whether LLMs have effectively acquired this knowledge. Specifically, knowledge probing is a technique that queries LLMs to assess their internal factual knowledge (Meng et al., 2022b). As illustrated in Figure 1, we probe LLMs with questions generated from biomedical knowledge triples to determine whether they can correctly recall the target knowledge. By comparing the knowledge probing results of LLMs before and after editing, we can evaluate how effectively knowledge editing enhances LLMs' ability to handle long-tail biomedical knowledge. Our key findings are:

- LLMs struggle to capture long-tail biomedical knowledge through pre-training.
- Knowledge editing can enhance LLMs' performance on long-tail biomedical knowledge, but it remains less effective compared to more common knowledge.
- Edited LLMs can memorise the form of long-tail knowledge, but their ability to generalise such knowledge is limited.
- We define one-to-many knowledge as triples where a single subject-relation pair is linked to multiple valid objects. This pattern is prevalent in long-tail biomedical knowledge and is a key factor leading to LLMs' poor performance in capturing long-tail knowledge.
- Effectively handling one-to-many knowledge is critical for improving LLMs' performance on long-tail biomedical knowledge through knowledge editing.

2 Background and Definitions

This section defines long-tail biomedical knowledge and briefly introduces the knowledge probing and editing techniques used in our experiments.

2.1 Long-Tail Biomedical Knowledge

We denote biomedical knowledge using knowledge triple $\langle s, r, o \rangle$, where s is the subject, r is the relation, and o is the object. Let $\mathcal{D}$ be the set of documents in the pre-training corpus, and $\mathcal{D}(s, o)$ be the subset of documents where both s and o co-occur. We define the *co-occurrence number* of the knowledge triple as $|\mathcal{D}(s, o)|$, which represents the frequency of knowledge $\langle s, r, o \rangle$ within the document set $\mathcal{D}$ (Kandpal et al., 2023). In this paper, following Mallen et al. (2023) and Kandpal et al. (2023), we define *long-tail knowledge* as:

$$\mathcal{K}_{\text{l}} = \{\langle s, r, o \rangle \mid |\mathcal{D}(s, o)| < \alpha\}, \quad (1)$$

where $\mathcal{K}_{\text{l}}$ denotes the set of long-tail knowledge and α represents a predefined threshold.

2.2 Knowledge Probing

Knowledge probing aims to evaluate LLMs' ability to capture factual knowledge (Meng et al., 2022b), and can serve as an evaluation method to assess the effectiveness of knowledge editing (Hernandez et al., 2023). Specifically, given a subject s and a relation r in a triple $\langle s, r, o \rangle$, we use a manually designed template $\mathcal{T}(s, r)$ to generate a natural language question, which is then fed into an LLM f_θ to generate the object o as the answer. Following prior works Meng et al. (2022a) and Kassner et al. (2021), accuracy (ACC) is commonly used to evaluate the performance of LLM in recalling the correct target entity o, which is formulated as:

$$\mathbb{E}_{\langle s,r,o \rangle \sim \mathcal{P}} \mathbb{I} \left\{ \arg\max_{y} f_\theta(y \mid \mathcal{T}(s, r)) = o \right\}, \quad (2)$$

where $\mathbb{E}_{\langle s,r,o \rangle \sim \mathcal{P}}$ denotes the expectation over a set of knowledge triples $\mathcal{P}$, y indicates the predicted answer and $\mathbb{I}\{\cdot\}$ is the indicator function. In this paper, we compare the knowledge probing results of LLMs before and after knowledge editing to investigate the effectiveness of editing methods in handling long-tail biomedical knowledge.

2.3 Knowledge Editing

Knowledge editing (Yao et al., 2023) aims to inject a new knowledge $\langle s, r, o \rangle$ into an LLM through a specific edit descriptor (x_e, y_e) (Yao et al., 2023). Given a knowledge $\langle s, r, o \rangle$ for editing, x_e can be formulated as $\langle s, r \rangle$, and $y_e = o$. The ultimate target of knowledge editing is to obtain an edited model f_{θ_e}, which effectively integrates the intended modifications within the editing scope, while preserving the model's performance for out-of-scope unrelated facts:

$$f_{\theta_e}(x) = \begin{cases} y_e & \text{if } x \in I(x_e, y_e) \\ f_\theta(x) & \text{if } x \in O(x_e, y_e) \end{cases} \quad (3)$$

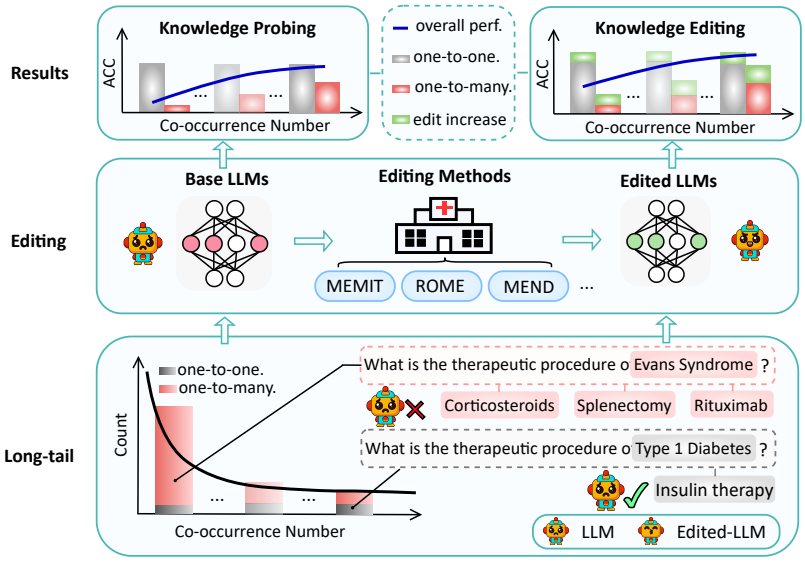


Figure 2: An overview of probing and editing for biomedical knowledge. These knowledge triples are classified into different groups based on co-occurrence number and further divided into one-to-one and one-to-many categories based on the number of correct answers (see § 4.4). The increasing performance with the number of co-occurrence number indicates that LLMs struggle to effectively capture long-tail biomedical knowledge before and after editing.

Here, the *in-scope* set $I(x_e, y_e)$ includes x_e and its equivalence neighborhood $N(x_e, y_e)$, which includes related input/output pairs. In contrast, the out-of-scope $O(x_e, y_e)$ contains inputs that are unrelated to the edit descriptor (x_e, y_e).

3 Identifying Long-Tail Biomedical Knowledge

Due to the lack of biomedical datasets specifically designed to evaluate long-tail knowledge, we develop a pipeline to extract such knowledge. In this section, we outline the procedures for extracting long-tail biomedical knowledge, with further details provided in Appendix A and Figure 7.

We focus on biomedical knowledge represented as knowledge triples and extract these triples from SNOMED CT (Donnelly et al., 2006), which is a comprehensive biomedical knowledge graph comprising over 1.4 million clinical triples (Benson and Grieve, 2021), and widely used to evaluate LLMs' understanding of biomedical knowledge (Meng et al., 2022b). Following previous work (Kandpal et al., 2023), we adopt the co-occurrence number—i.e., how often a triple's subject and object appear in the same document—as a proxy for knowledge popularity. To identify the long-tail knowledge within these triples, we use an entity linking pipeline to compute the co-occurrence number of each triple in the PubMed corpus[2], which is a widely used biomedical corpus for pre-training. In the entity linking pipeline, we use PubTator (Wei et al., 2013) to annotate entities in the PubMed corpus and then use SapBERT (Liu et al., 2021) to link knowledge triple entities to PubMed entities. Subsequently, we compute the co-occurrence number for each triple. Long-tail knowledge is defined as triples with a co-occurrence number less than 10 (Kandpal et al., 2023). As a result, we obtained 59,705, 14,087, and 28,375 triples for the training, validation, and test sets, respectively, stratified by varying levels of co-occurrence. The statistics of the dataset are presented in Table 1. We refer to our dataset as **CliKT** (Clinical Knowledge Triples).

To evaluate LLMs' ability to understand these triples, we generate question-answer pairs following Meng et al. (2022a). For each triple, we construct a question using the subject and relation, with the object serving as the answer. For example, for the triple $\langle$*Diabetes, treated_by, Insulin*$\rangle$, the corresponding QA pair is: *What is Diabetes treated by? Answer: Insulin*. The template for constructing questions is provided in Table 3.

4 Knowledge Editing for Long-Tail Biomedical Knowledge

In this section, we investigate the effectiveness of knowledge editing methods in enhancing LLMs' ability to handle long-tail biomedical knowledge. Since some editing methods, e.g., MEND (Mitchell et al., 2022) and IKE (Zheng et al., 2023a), require

[2] https://pubmed.ncbi.nlm.nih.gov/

Item	Train	Valid	Test
# Triples	59,705	14,087	28,375
$\lvert\mathcal{D}(s,o)\rvert < 10^1$	52,297	11,476	22,952
$\lvert\mathcal{D}(s,o)\rvert \in [10^1, 10^2)$	5,363	2,055	4,110
$\lvert\mathcal{D}(s,o)\rvert \in [10^2, 10^3)$	1,659	551	1,103
$\lvert\mathcal{D}(s,o)\rvert \geq 10^3$	386	105	210
# Relations	21	21	21
# Subjects	39,654	12,267	21,872
# Objects	7,867	3,526	4,706

Table 1: The statistics of CliKT dataset. $|\mathcal{D}(s,o)|$ represents the oc-occurrence number of knowledge triple.

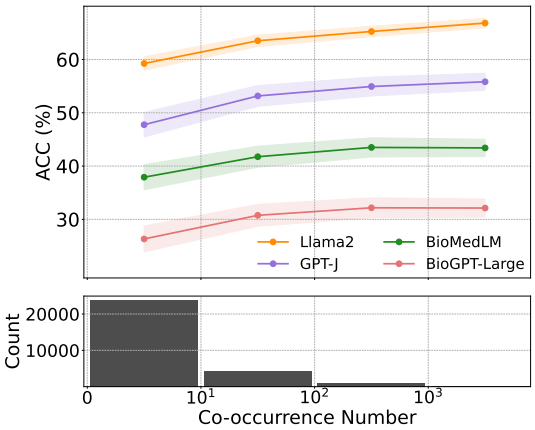


Figure 3: The overall performance of pre-edit probing on Llama2, GPT-J, BioMedLM and BioGPT-Large. The shaded areas indicate the standard deviation and Count denotes the number of triples within each group.

training data, we follow the data splitting strategy proposed by Meng et al. (2022a) to divide our CliKT dataset into training, validation, and test sets (see Table 1)[3]. We report all results on the test set.

4.1 Experimental Setup

LLMs. To investigate whether LLMs can be edited for long-tail biomedical knowledge, we focus on LLMs that are specifically pre-trained on biomedical data. We employ two models primarily trained on PubMed: **BioGPT-Large** (Luo et al., 2022) and **BioMedLM** (Bolton et al., 2024). Furthermore, we include four general-domain LLMs: **Llama2** (Touvron et al., 2023), **Llama3** (Grattafiori et al., 2024), **GPT-J** (Wang and Komatsuzaki, 2021) and **Qwen2.5** (Yang et al., 2024) to evaluate whether our findings generalise to models not specifically trained on biomedical data[4].

Knowledge Editing Methods. For knowledge editing, we employ the following methods, which have demonstrated strong effectiveness in knowledge injection tasks (Wang et al., 2025):

- **ROME** (Meng et al., 2022a): ROME updates an MLP layer to encode new information by treating the MLP module as a key-value memory. It relies on causal mediation analysis to precisely identify the location for editing.
- **MEMIT** (Meng et al., 2023): it employs the localisation strategies from ROME and applies explicit parameter adjustments to inject new knowledge across multiple layers.
- **MEND** (Mitchell et al., 2022): MEND enables efficient, targeted updates to LLMs by leveraging low-rank gradient transformations. It enables quick, localised modifications in model behaviour using only a single input-output example, while preventing overfitting.
- **IKE** (Zheng et al., 2023a): IKE modifies factual knowledge in LLMs through in-context learning without updating parameters. It corrects specific knowledge using demonstration contexts, reducing over-editing and preserving previously stored knowledge.
- **FT** (Yao et al., 2023): FT updates model parameters using gradient descent on a single MLP layer identified by ROME. We employ the FT implementation within the EasyEdit framework (Wang et al., 2023b).

We follow the official implementations for each method and perform hyperparameter tuning on our CliKT dataset to ensure a fair comparison[5].

Evaluation Metrics. We use knowledge probing to assess whether LLMs have successfully acquired biomedical knowledge within the CliKT dataset. Specifically, we assess their zero-shot QA performance on the test-set questions, using accuracy (**ACC**) as the evaluation metric, as detailed in § 2.2.

In addition, we adopt standard knowledge editing metrics (Meng et al., 2022a; Yao et al., 2023) to assess the effectiveness of editing: (1) **Reliability** measures whether the model correctly incorporates the target knowledge after editing—i.e., whether it outputs the correct answer for the edited input; (2) **Generalisation** evaluates whether the model can apply the updated knowledge to semantically similar variations (e.g., paraphrased queries), reflecting the robustness of the edit; (3) **Locality** assesses whether unrelated predictions remain unaffected

[3]Details of dataset splitting method are in Appendix A.3.
[4]Details of these LLMs are provided in Appendix B.1.
[5]Details about the training and hyperparameter tuning process can be found in Appendix B.4.

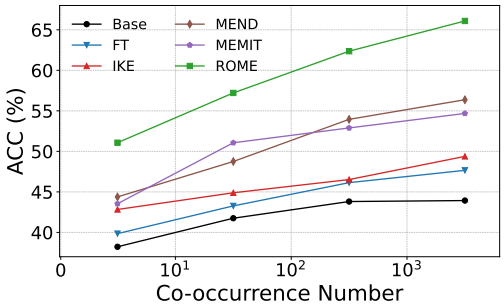


Figure 4: The performance of knowledge probing after editing with different editing methods on BioMedLM, where "Base" denotes LLM without editing.

after editing, ensuring that edits are localized and do not introduce unintended side effects.

Evaluation examples for these three metrics are derived from the test set of CliKT. Due to space limit, more details about metric definitions, evaluation example construction procedures and illustrative examples are provided in Appendix B.2.

4.2 Pre-Edit Results on Long-Tail Biomedical Knowledge

Finding 1: *LLMs struggle to capture long-tail biomedical knowledge through pre-training.*

To investigate whether LLMs face challenges in capturing long-tail biomedical knowledge during pre-training, we categorise biomedical knowledge triples in CliKT into different groups based on their co-occurrence number $|\mathcal{D}(s, o)|$ and evaluate the probing results of LLMs across these groups.

The bottom portion of Figure 3 shows the distribution of triples across the different groups, which highlights the long-tail nature of biomedical knowledge, where long-tail knowledge accounts for the majority of the data. The results for biomedical LLMs and general-domain LLMs are illustrated in the top portion of Figure 3. Specifically, Figure 3 shows that the performance of LLMs declines as the co-occurrence number decreases. In particular, the performance of BioMedLM on long-tail knowledge ($|\mathcal{D}(s, o)| < 10$) is 22.86% lower relative to its performance on popular knowledge ($|\mathcal{D}(s, o)| \geq 10^3$). This trend is also evident in general-domain LLMs. For example, Llama2 experiences an accuracy drop of 16.86% when handling long-tail biomedical knowledge compared with popular knowledge. These results indicate that LLMs struggle with long-tail biomedical knowledge, highlighting the challenge of accurately capturing long-tail knowledge during pre-training. Furthermore, Figure 3 shows that as the co-occurrence number decreases, the standard deviation of ACC increases. This observation implies that LLMs exhibit greater confidence when processing popular biomedical knowledge than long-tail biomedical knowledge.

Group	Edit	Reliability↑	Gen.↑	Locality↑
$<10^1$	ROME	**98.02**	**68.42**	83.70
	MEMIT	86.21	47.36	**98.10**
	MEND	91.32	46.75	89.60
	IKE	83.87	43.70	97.81
	FT	32.52	40.36	96.80
$[10^1, 10^2)$	ROME	**98.11**	**70.10**	84.60
	MEMIT	89.21	48.21	97.30
	MEND	88.90	47.80	89.83
	IKE	84.52	45.12	96.80
	FT	33.35	40.78	**97.90**
$[10^2, 10^3)$	ROME	**98.63**	**72.50**	84.62
	MEMIT	89.01	51.47	**97.90**
	MEND	88.94	48.83	91.40
	IKE	85.89	46.74	96.85
	FT	33.89	44.62	96.66
$\geq 10^3$	ROME	**98.66**	**72.54**	84.45
	MEMIT	89.87	50.00	97.43
	MEND	90.96	49.86	90.92
	IKE	85.91	48.76	96.87
	FT	34.84	44.62	**97.57**

Table 2: Performance of knowledge editing methods on the CliKT dataset across different co-occurrence number groups. The best performance per group is marked in boldface, while the second-best performance is underlined. ↑ indicates that higher values reflect better performance, and "Gen." stands for Generalisation.

Based on the above analysis, we conclude that LLMs indeed struggle to capture long-tail biomedical knowledge. As long-tail knowledge constitutes the majority of biomedical data, it is crucial to explore methods that can effectively improve LLMs' performance on long-tail biomedical knowledge.

4.3 Post-Edit Results for Long-Tail Biomedical Knowledge

Finding 2: *Knowledge editing can enhance LLMs' performance on long-tail biomedical knowledge, but it remains less effective compared to more common knowledge.*

Subsequently, we investigate the effectiveness of knowledge editing for long-tail biomedical knowledge. We apply existing knowledge editing methods to inject biomedical knowledge from the CliKT dataset into LLMs and then follow the procedures in the pre-edit experiments for evaluation.

The post-edit probing results for BioMedLM[6] are presented in Figure 4. These results yield the

[6]The results of other LLMs, i.e., BioGPT, Llama2, Llama3, Qwen2.5, can be found in Figure 8 of the Appendix, which show similar findings as BioMedLM.

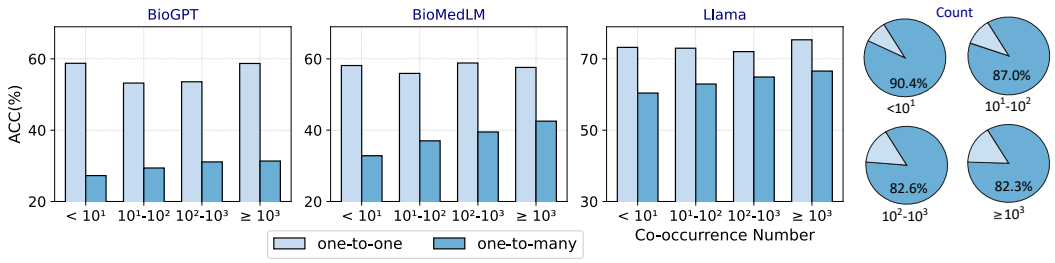


Figure 5: The comparison of knowledge probing performance between one-to-one and one-to-many settings across different co-occurrence numbers, with the pie chart on the far right illustrating the data distribution.

following findings: (1) Knowledge editing methods, especially ROME, can enhance LLM's ability in handling long-tail biomedical knowledge. For example, Figure 4 shows that BioMedLM edited with ROME achieves an improvement of approximately 52.08% in ACC on long-tail knowledge ($|\mathcal{D}(s,o)| < 10$) compared to the base model before editing; (2) Despite the improvements from knowledge editing, Figure 4 also reveals that ACC of post-edit LLMs consistently drops as the co-occurrence number decreases across all the editing methods. Specifically, for ROME, the ACC on long-tail knowledge is still 16.15% relatively lower than on popular knowledge ($|\mathcal{D}(s,o)| \geq 10^3$). This indicates that even after editing, the edited LLMs still struggle with long-tail knowledge.

Finding 3: *Edited LLMs can memorise the form of long-tail knowledge, but their ability to generalise such knowledge is limited.*

In addition to the post-edit probing results, we also calculate the other editing metrics outlined in §4.1 to comprehensively evaluate the effectiveness of the editing methods. Specifically, we calculate the Reliability, Generalisation and Locality metrics of edited models across different groups of biomedical knowledge. From the results in Table 2, we observe that ROME's Reliability remains above 98% across all groups, with no significant variation. Similarly, the Reliability of MEMIT, MEND, and IKE is largely unaffected by the co-occurrence number, indicating that the edited LLMs' ability to memorise the form of inserted knowledge is not influenced by long-tail knowledge. However, the generalisation performance declines as the co-occurrence number decreases, which aligns with the observed reduction in post-edit ACC for edited-LLMs as the co-occurrence number decreases. This observation suggests that, although edited LLMs can memorize the form of long-tail knowledge itself after knowledge editing, their ability to generalise this long-tail knowledge, especially in reasoning and responding to related questions, remains influenced by low co-occurrence numbers.

Furthermore, we observe that, though all the editing methods exhibit relatively strong performance in terms of locality across groups, ROME is affected more than the other methods. This indicates that while ROME achieves the best reliability and generalisation, it may slightly affect unrelated knowledge, consistent with the observations of Wang et al. (Wang et al., 2024b).

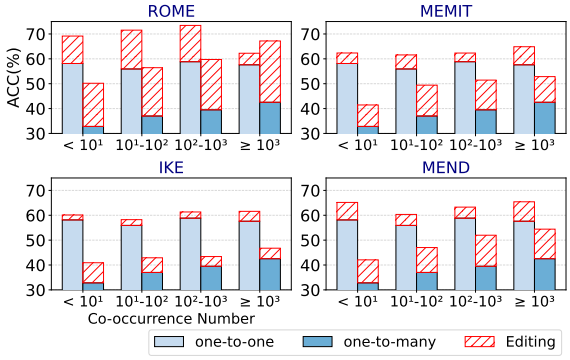


Figure 6: The knowledge probing performance of BioMedLM on both one-to-one knowledge and one-to-many knowledge before and after editing.

4.4 Knowledge Type Analysis in Editing

In this section, to further investigate the cause of the performance gap between long-tail and popular biomedical knowledge before and after editing, we further subdivide both long-tail and popular knowledge into two categories: *one-to-one* and *one-to-many*. The *one-to-one* knowledge refers to triples where a subject is linked to a single object via a given relation, while *one-to-many* knowledge represents triples where the same subject-relation pair is linked to multiple objects (Nagasawa et al., 2023). For example, the triple ⟨*Type 1 diabetes, therapeutic procedure, insulin therapy*⟩ represents a one-to-one knowledge, where "Type 1 diabetes" is associated with a single object, "insulin therapy".

In contrast, ⟨*hypertension, associated with, heart disease*⟩ exemplifies a one-to-many knowledge, where "hypertension" can be linked to multiple objects, such as "stroke" or "kidney disease"[7].

4.4.1 Pre-Edit Probing of Different Types of Knowledge

Finding 4: *The prevalence of one-to-many knowledge in long-tail biomedical knowledge is a key factor contributing to LLMs' poor performance in capturing such long-tail knowledge*.

Figure 5 shows the pre-edit probing results of one-to-one and one-to-many knowledge across different co-occurrence number groups. We found that one-to-one knowledge is almost unaffected by co-occurrence numbers and consistently outperforms one-to-many knowledge in all groups. For instance, BioGPT achieves an ACC that is approximately 115.56% higher on one-to-one knowledge compared to one-to-many knowledge. In contrast, for one-to-many knowledge, results from BioGPT, BioMedLM, and Llama2 all show a steady increase in ACC as the co-occurrence number increases. This suggests that co-occurrence number, or knowledge frequency, has a significant impact on LLMs' ability to accurately comprehend one-to-many knowledge. We further analysed the distribution of one-to-one and one-to-many knowledge. Figure 5 shows that as the co-occurrence number increases, the proportion of one-to-many knowledge decreases while one-to-one knowledge increases. In the long-tail knowledge group ($|\mathcal{D}(s,o)| < 10$), 90.4% of the knowledge is one-to-many. This analysis reveals that LLMs' difficulty with long-tail biomedical knowledge before editing is primarily due to the large proportion of one-to-many knowledge, which is challenging for LLMs to comprehend, as it increases the probability that the correct answers will not align with the model's output.

4.4.2 Knowledge Editing for Different Types of Knowledge

Finding 5: *Effectively handling one-to-many knowledge is critical for improving LLMs' performance on long-tail biomedical knowledge through knowledge editing*.

Next, we apply editing methods to both one-to-one and one-to-many knowledge. The results for BioMedLM[8] are provided in Figure 6, which indicate that while editing methods enhance performance on one-to-many knowledge, the improvement remains limited. For instance, in the ROME-edited BioMedLM for the long-tail knowledge ($|\mathcal{D}(s,o)| < 10$), the ACC for one-to-one knowledge was initially 42.19% higher than that for one-to-many knowledge. After applying the editing, this gap decreased to 16.43%. However, the persistent gap also highlights that even after editing, the model's performance on one-to-many knowledge, which constitutes the majority of long-tail knowledge, remains constrained. This finding suggests that *despite knowledge editing can enhance LLMs' capability in handling one-to-many knowledge, there remains a challenge in bridging the performance gap between one-to-one and one-to-many knowledge*. This limitation is critical given that one-to-many knowledge constitutes the majority of long-tail knowledge.

5 Related Work

5.1 LLMs for the Biomedical Domain

LLMs have achieved remarkable progress in the biomedical domain (Tian et al., 2024). Early advance were led by BERT (Vaswani et al., 2017) and its variants, such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019), which showed significant improvements in named entity recognition and relation extraction when applied to large datasets such as PubMed and clinical notes (Perera et al., 2020; Sun et al., 2021). GPT-based models, including GPT-J (Wang and Komatsuzaki, 2021), BioGPT (Luo et al., 2022) and BioMedLM (Bolton et al., 2024), further enhanced biomedical text generation and question answering (Tian et al., 2024). Recent LLMs like Llama (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), and Palm (Chowdhery et al., 2023) have scaled transformer architectures to address more complex tasks, such as biomedical knowledge reasoning (Wu et al., 2024a; Watanabe et al., 2024) and assisting in clinical decision-making (Sandmann et al., 2024). This work explores LLMs' performance on long-tail biomedical knowledge. We present the first study to investigate how long-tail knowledge impacts LLMs in knowledge editing, offering new insights into improving

[7]The detailed evaluation process of one-to-one and one-to-many knowledge, following the same procedure described in Section 4.1, can be found in Appendix B.3.

[8]The results for other LLMs, i.e., BioGPT, Llama2, Llama3, Qwen2.5, are provided in Figure 9 and Figure 10, which demonstrate similar results as BioMedLM.

LLMs' handling of rare biomedical information through knowledge editing techniques.

5.2 Knowledge Editing

Existing knowledge editing methods can be classified into three categories (Yao et al., 2023): memory-based (Zheng et al., 2023b), meta learning (Mitchell et al., 2022), and locate-then-edit (Meng et al., 2022a). Memory-based methods, like IKE (Zheng et al., 2023b), leverage external memory to update knowledge without changing model parameters. Meta-learning methods, such as KE (Cao et al., 2021), train a hyper-network to generate updated weights. MEND (Mitchell et al., 2022) improves on this by using low-rank gradient updates for more efficient model edits.

Locate-then-edit approaches aim for more targeted knowledge editing. Methods like KN (Dai et al., 2022) use knowledge attribution to locate relevant neurons but struggle with precise weight updates. ROME (Meng et al., 2022a) advances this by using causal tracing to locate and edit the Feed Forward Network (FFN) layers, which act as key-value memories (Geva et al., 2021, 2023). MEMIT (Meng et al., 2023) further expands this technique for batch editing. To the best of our knowledge, this work is the first to investigate the effectiveness of knowledge editing on long-tail biomedical knowledge.

5.3 Long-Tail Knowledge within LLMs

Existing studies have explored how long-tail knowledge, affects LLMs' performance (Shin et al., 2022; Han and Tsvetkov, 2022; Elazar et al., 2022; Mallen et al., 2023; Kandpal et al., 2023). Mallen et al. (2023) find that commonsense QA accuracy is strongly correlated with the frequency of entity popularity in the pre-training data from Wikipedia (Milne and Witten, 2008). Similarly, Elazar et al. (2022) employ causal inference to investigate how pre-training data statistics affect commonsense QA, highlighting how models rely on co-occurrence patterns between subjects, objects, and text to answer questions. More recently, Kandpal et al. (2023) explore the connection between the knowledge LLMs acquire for general-domain QA tasks and its frequency in the pre-training corpus, introducing comparative experiments involving model retraining and scaling.

Despite these findings, most prior works have focused on general-domain QA, leaving the long-tail biomedical domain remaining largely unexplored (Wu et al., 2024b). This gap is especially concerning as LLMs are increasingly being used by healthcare professionals, including doctors, to assist in diagnosis and treatment recommendations. Our research fills this gap by investigating the influence of long-tail biomedical knowledge on LLMs through knowledge probing and examining its impact on the effectiveness of knowledge editing.

6 Discussion

While our work highlights the challenges LLMs face in capturing and editing biomedical one-to-many knowledge, we acknowledge that addressing these limitations requires further exploration. We outline several promising directions that may help improve performance in this domain: (1) Retrieval-augmented generation (RAG): incorporating external biomedical knowledge by retrieving relevant documents or triples could help LLMs better handle long-tail biomedical knowledge. This approach has shown promise in open-domain QA (Gao et al., 2023) and may be adapted for biomedical editing with domain-specific retrieval modules; (2) Structure-aware finetuning: instead of treating each triple independently, future work could explore fine-tuning strategies that explicitly model the structure of one-to-many knowledge. For example, training objectives can be designed to encourage the model to recognise that multiple objects may be valid for a given subject-relation pair.

7 Conclusion

In this paper, we investigate the effectiveness of knowledge editing methods for addressing the challenges of long-tail biomedical knowledge in LLMs. Our results show that while existing techniques enhance performance on long-tail knowledge, they still fall short compared to their performance on high-frequency knowledge. This disparity is largely due to the prevalence of one-to-many knowledge structures in the biomedical domain, which complicate models' ability to accurately represent and edit such information. Our results highlight the need for advanced editing techniques specifically designed for long-tail knowledge. These techniques should prioritise strategies for effectively handling the intricacies of one-to-many knowledge scenarios, which are particularly common in the biomedical domain and remain a significant obstacle for current methods.

Limitations

We identify the following limitations of our work: (1) First, our approach to extracting long-tail knowledge is based on document-level co-occurrence frequency (Kandpal et al., 2023), which captures general patterns of occurrence but lacks refinement at the sentence level. This limitation may cause our analysis to miss finer patterns in knowledge distribution, especially in instances where sentence-level context provides essential nuances. Future work could enhance the long-tail knowledge extraction pipeline by investigating co-occurrence on the sentence-level to improve the granularity of knowledge editing. (2) Second, our experimental framework is limited to the collection of over 100,000 biomedical knowledge extracted from PubMed, an extensive repository of biomedical literature. While we believe the scale of this collection offers a robust foundation for evaluating our methods, our future research should focus on extracting long-tail knowledge from a broader range of domains to further validate the generalisability of our findings. (3) Finally, we concentrate on analysing limitations without proposing specific solutions, prioritising the establishment of a comprehensive understanding. Future work will focus on developing methods to improve knowledge editing performance on long-tail knowledge.

References

Mohd Hafizul Afifi Abdullah, Norshakirah Aziz, Said Jadid Abdulkadir, Hitham Seddig Alhassan Alhussian, and Noureen Talpur. 2023. Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges. *IEEE Access*, 11:10535–10562.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Israa Alghanmi, Luis Espinosa Anke, and Steven Schockaert. 2021. Probing pre-trained language models for disease knowledge. In *Findings of the Association for Computational Linguistics*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3023–3033.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Tim Benson and Grahame Grieve. 2021. *SNOMED CT*, pages 293–324. Springer International Publishing, Cham.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and 1 others. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*, pages 8493–8502.

Julien Delile, Srayanta Mukherjee, Anton Van Pamel, and Leonid Zhukov. 2024. Graph-based retriever captures the long tail of biomedical knowledge. *arXiv preprint arXiv:2402.12352*.

Kevin Donnelly and 1 others. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model'sfactual'predictions. *arXiv preprint arXiv:2207.14251*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiaochuang Han and Yulia Tsvetkov. 2022. Orca: Interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *arXiv preprint arXiv:2205.12600*.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

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*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3250–3258.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18564–18572.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4228–4238.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9802–9822.

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. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022b. Rewire-then-Probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4798–4810.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

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

Haruki Nagasawa, Benjamin Heinzerling, Kazuma Kokuta, and Kentaro Inui. 2023. Can lms store and retrieve 1-to-n relational knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 130–138.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *Proceedings of the Web Conference 2021*, pages 2636–2647.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673.

Richard J Roberts. 2001. Pubmed central: The genbank of the published literature.

Sarah Sandmann, Sarah Riepenhausen, Lucas Plagwitz, and Julian Varghese. 2024. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050.

Pranav Shetty and Rampi Ramprasad. 2021. Automated knowledge extraction from polymer literature using natural language processing. *Iscience*, 24(1).

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woo-Myoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5168–5186.

Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2021. Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 118:103799.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, and 1 others. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023a. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-lingual knowledge editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11676–11686.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024b. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *arXiv preprint arXiv:2405.14768*.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, and 1 others. 2023b. 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, and Jundong Li. 2025. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3):59:1–59:37.

Natsumi Watanabe, Kudoro Kinasaka, and Akira Nakamura. 2024. Empower llama 2 for advanced logical reasoning in natural language understanding.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Zheng Wu, Kehua Guo, Entao Luo, Tian Wang, Shoujin Wang, Yi Yang, Xiangyuan Zhu, and Rui Ding. 2024b. Medical long-tailed learning for imbalanced data: bibliometric analysis. *Computer Methods and Programs in Biomedicine*, page 108106.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023a. 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.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023b.

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.

Appendix

In the Appendix, we introduce more details along with dataset construction, experimental details, and additional experimental results:

- **Appendix A**: CliKT Construction (cf. Section 3).
- **Appendix B**: Experimental Details (cf. Section 2 and 3).
- **Appendix C**: Additional Results (cf. Section 3).

A CliKT Construction

Due to the lack of datasets dedicated to evaluating long-tail biomedical knowledge, we propose CliKT, a new benchmark specifically designed to evaluate LLMs' performance on long-tail biomedical knowledge. Notably, given that PubMed is a widely used biomedical corpus for pre-training LLMs (Wang et al., 2023a), which contains over 37 million abstracts of biomedical papers (Wei et al., 2013), we mainly focus on PubMed data to extract long-tail biomedical knowledge. Specifically, we first extract knowledge triples from SNOMED CT (Donnelly et al., 2006) (§A.1) to obtain a comprehensive set of biomedical concepts and their relationships. Next, we employ an entity linking pipeline to map these triples back to their corresponding documents in the PubMed (Roberts, 2001) corpus (§A.2), enabling us to identify whether a triple represents long-tail knowledge based its occurrence in the corpus. Finally, we generate question-answer (QA) pairs based on the knowledge triples to evaluate the ability of LLMs to capture the factual knowledge, and conduct a human evaluation to show that our entity linking pipeline accurately identifies relevant documents for the majority of the QA pairs.

A.1 Extracting Biomedical Knowledge Triples

We focus on the long-tail biomedical knowledge from the PubMed corpus. However, directly extracting such knowledge from the entire corpus is a challenging task (Shetty and Ramprasad, 2021; Nguyen et al., 2021; Abdullah et al., 2023). Therefore, following previous work (Alghanmi et al., 2021; Fei et al., 2021), we leverage information from existing biomedical knowledge graphs to facilitate more efficient extraction. Specifically, we extract all the knowledge triples from SNOMED CT (Donnelly et al., 2006), which is a comprehensive biomedical knowledge graph comprising over 200K triples and widely used for assessing LLMs' understanding of biomedical knowledge (Meng et al., 2022b). Each triple is denoted as (head entity, relation, tail entity), representing the relationship between two entities, e.g., (Type 1 Diabetes, Therapeutic Procedure, Insulin therapy).

A.2 Mapping Knowledge Triples to PubMed Documents

We then develop an entity linking pipeline to map the extracted knowledge triples back to documents in Pubmed (Roberts, 2001) to identify long-tail knowledge. The detailed procedure is as follows:

Entity Annotation. To facilitate the mapping of knowledge triples to specific PubMed documents, we first need to annotate the entities within the PubMed corpus. To this end, we use PubTator (Wei et al., 2013), a robust web-based text-mining tool that provides automatic annotations of biomedical concepts in PubMed. Following the work of Wei et al. (2019), we obtain entity annotations within 37 million PubMed abstracts[9].

Entity Linking. After obtaining annotated entities, the next step is to map the knowledge triples to their corresponding PubMed documents. Previous studies (Elsahar et al., 2018; Kandpal et al., 2023) suggest that when the head entity and the tail entity of a knowledge triple co-occur within a document, it is likely that the knowledge represented by the triple is expressed in that document. Based on this observation, we define documents where both the head and tail entities of a knowledge triple co-occur as its *related documents*, and the count of such documents as the *co-occurrence number*.

[9]The annotated data is available at https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/

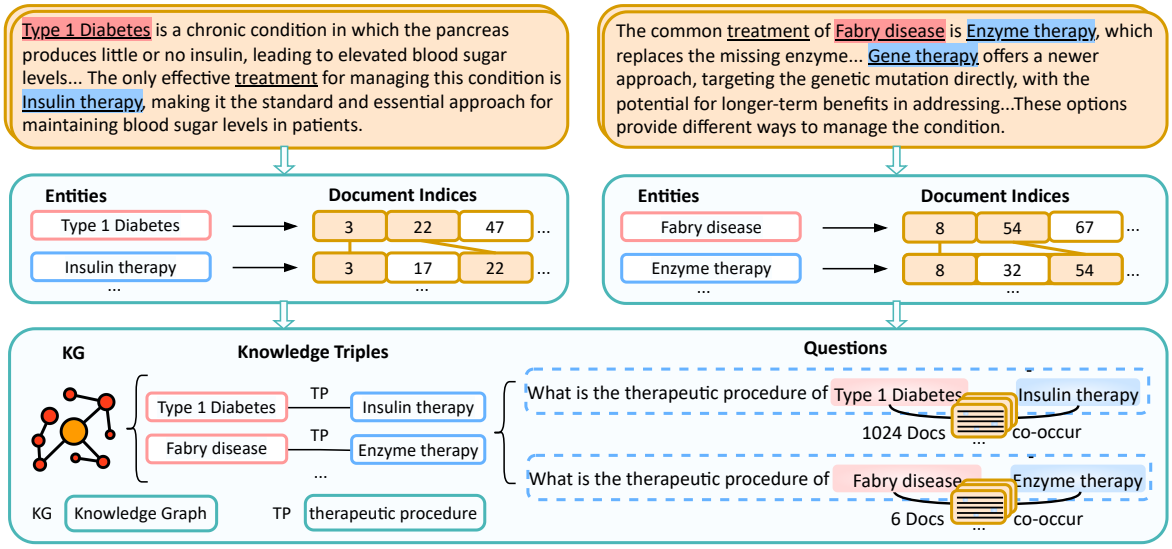


Figure 7: The pipeline for identifying long-tail biomedical knowledge consists of a systematic process encompassing document collection, entity linking, knowledge graph traversal, and question generation.

To determine whether both the head and tail entities of a triple co-occur in a document, we use SapBERT (Liu et al., 2021), an effective biomedical entity linking model, to match these entities to those present in the document. For instance, given the triple (Hypertension, causes, heart disease) from SNOMED CT, SapBERT can link "Hypertension" to its equivalent term "high blood pressure" in PubMed, ensuring an accurate match with related documents. We iterate through the entire corpus to calculate the co-occurrence number for each triple. We define triples with a low co-occurrence number as long-tail biomedical knowledge.

Question Generation. Finally, we generate QA pairs based on the resulting triples to assess the LLMs' ability to capture these knowledge triples. Following Meng et al. (2022a), we manually design templates to generate questions using the head entity and the relation, while considering the tail entity as the answer. For example, given a triple (Diabetes, treated_by, Insulin), the corresponding QA pair would be: *Question: What is Diabetes treated by? Answer: Insulin*. We provide some example templates in Table 3, where "**Question**" is the template used for constructing questions.

A.3 Dataset Splitting

After generating the question-answer pairs, we randomly split them into training, validation and test sets using an 7:1:2 ratio. Following the initial split, we applied additional filtering to the training set by discarding knowledge triples with zero co-occurrence number, resulting in a slightly smaller effective training set. The detailed statistics of each split are provided in Table 1. To preserve the natural distribution and diversity of relational patterns, we did not explicitly constrain the overlap of subjects or objects across splits. As a result, some entities may appear in multiple sets. This design choice ensures a realistic and challenging setting for evaluating editing methods that may rely on generalisation across related facts.

B Experimental Details

B.1 Details of Large Language Models

We employ two biomedical LLMs and two general-domain LLMs in our experiments:

- **BioGPT-Large (Luo et al., 2022):** A 1.5 billion parameter model from Microsoft, primarily pre-trained on PubMed, excelling in drug discovery and medical record analysis.
- **BioMedLM (Bolton et al., 2024):** A Stanford-developed model optimised for biomedical tasks, pretrained on PubMed with 2.7 billion parameters, ideal for literature retrieval and information extraction.

Relation	Template
Finding site	**Edit Prompt:** "The finding site of [SUBJECT] is." **Question:** "What is the finding site of [SUBJECT]?" **Rephrase:** "Where is [SUBJECT] typically found?"
Associated morphology	**Edit Prompt:** "The associated morphology of [SUBJECT] is." **Question:** "What is the associated morphology of SUBJECT?" **Rephrase:** "Can you describe the morphology associated with [SUBJECT]"
Causative agent	**Edit Prompt:** "The causative agent of [SUBJECT] is" **Question:** "What is the causative agent of [SUBJECT]?" **Rephrase:** "Which pathogen causes [SUBJECT]?"
Interprets	**Edit Prompt:** "[SUBJECT] interprets." **Question:** "What does [SUBJECT] interprets?" **Rephrase:** "What is interpreted by [SUBJECT]?"
Procedure site	**Edit Prompt:** "The procedure site of [SUBJECT] is" **Question:** "What is the indirect procedure site of [SUBJECT]?" **Rephrase:** "Where is the procedure site for [SUBJECT]?"
Pathological process	**Edit Prompt:** "The pathological process of [SUBJECT] involves." **Question:** "What is the pathological process of [SUBJECT]?" **Rephrase:** "Which pathological process does [SUBJECT] involve?"
Due to	**Edit Prompt:** "[SUBJECT] is due to." **Question:** "What is the [SUBJECT] due to?" **Rephrase:** "What is the cause of [SUBJECT]?"
Has active ingredient	**Edit Prompt:** "The active ingredient of [SUBJECT] is." **Question:** "What is the active ingredient of [SUBJECT]?" **Rephrase:** "What active ingredient does [SUBJECT] have?"
Part of	**Edit Prompt:** "[SUBJECT] is a part of." **Question:** "What is the [SUBJECT] a part of?" **Rephrase:** "To what is [SUBJECT] a part?"
Has definitional manifestation	**Edit Prompt:** "The definitional manifestation of [SUBJECT] is." **Question:** "What is the definitional manifestation of [SUBJECT]?" **Rephrase:** "How is [SUBJECT] manifested definitionally?"
Component	**Edit Prompt:** "The component of [SUBJECT] is." **Question:** "What is the component of [SUBJECT]?" **Rephrase:** "What components does [SUBJECT] consist of?"

Table 3: Examples of relation templates demonstrate how each relation is transformed into input prompts, which can categorized into three parts: Edit Prompt, Question, and Rephrase. The "Edit Prompt" is used for knowledge editing and reliability evaluation, the "Question" is designed for knowledge probing, and the "Rephrase" is used to assess generalisation metrics. The complete template for all the relations can be found in our github repository.

- **Llama2 (Touvron et al., 2023):** A Meta-developed model with 7 billion parameters, designed for general-purpose language tasks. It has been leveraging large-scale pretraining on diverse datasets, including biomedical corpora.
- **GPT-J (Wang and Komatsuzaki, 2021):** A 6 billion parameter open-source model by EleutherAI, trained on the Pile dataset, which includes a significant portion of biomedical texts from PubMed.

In addition to the models listed above, we also include results for two recently released models, Llama3 (Grattafiori et al., 2024) and Qwen2.5 (Yang et al., 2024), to provide a broader view of knowledge editing performance across both biomedical-specific and general-purpose LLMs.

B.2 Details of Knowledge Editing Evaluation Metrics

To evaluate the effectiveness of knowledge editing, we adopt three standard metrics: Reliability, Generalisation, and Locality. All evaluation instances are derived from the test split of the CliKT dataset. Below, we define each metric, describe how its evaluation data is constructed, and provide illustrative examples.

(1) **Reliability**: This metric evaluates whether the model has correctly incorporated target knowledge after editing. Specifically, it measures the model's accuracy on a set of test instances (x_e, y_e) that directly

correspond to the target edits.

$$\mathbb{E}_{x'_e, y'_e \sim \{(x_e, y_e)\}} \mathbf{1} \left\{ \underset{y}{\operatorname{argmax}} f_{\theta_e}(y \mid x'_e) = y'_e \right\}. \tag{4}$$

Construction Procedure. For each knowledge triple we aim to edit, e.g., (*Type 1 Diabetes, Therapeutic Procedure, Corticosteroids*), we first use an **Edit Prompt**, such as "*The therapeutic procedure of Type 1 Diabetes is Corticosteroids.*" to inject the knowledge into the model. We then use a corresponding evaluation question, such as "*What is the therapeutic procedure of Type 1 Diabetes?*", paired with its correct answer "*Corticosteroids*", to assess whether the edit was successful. These input-output pairs form the test set used to compute the reliability score.

(2) **Generalisation**: Considering that paraphrased sentences are modified accordingly through editing, this metric measures the average accuracy on equivalent neighbours $R(x_e, y_e)$, where equivalent neighbours are rephrased questions based on the edited knowledge. This metric evaluates the model's ability to apply the edited knowledge to semantically equivalent but surface-form-different inputs. It reflects whether the edit generalises beyond the exact phrasing used during editing. Formally, it measures the accuracy on a set of paraphrased input-output pairs $R(x_e, y_e)$:

$$\mathbb{E}_{x'_e, y'_e \sim R(x_e, y_e)} \mathbf{1} \left\{ \underset{y}{\operatorname{argmax}} f_{\theta_e}(y \mid x'_e) = y'_e \right\}. \tag{5}$$

Construction Procedure. Given a factual triple targeted for editing, e.g., (*Type 1 Diabetes, Therapeutic Procedure, Corticosteroids*), we first construct an evaluation question in canonical form, such as "*What is the therapeutic procedure of Type 1 Diabetes?*". To assess generalisation, we generate another semantically equivalent paraphrases of this question, e.g., "*Which treatment is used for Type 1 Diabetes?*" or "*How is Type 1 Diabetes typically treated?*". These paraphrases are created using a predefined **Rephrase** template. The expected answer "Corticosteroids" remains unchanged across all variants, and the model's ability to produce the correct answer across paraphrases indicates the strength of generalisation.

(3) **Locality**: This metric assesses whether the knowledge edit remains localized—that is, whether the model's behavior on unrelated inputs remains unchanged after editing. It reflects the extent to which the edit introduces undesired side effects on out-of-scope content. Formally, locality measures the consistency between the model's pre-edit and post-edit predictions over a set of unrelated input examples $O(x_e, y_e)$.

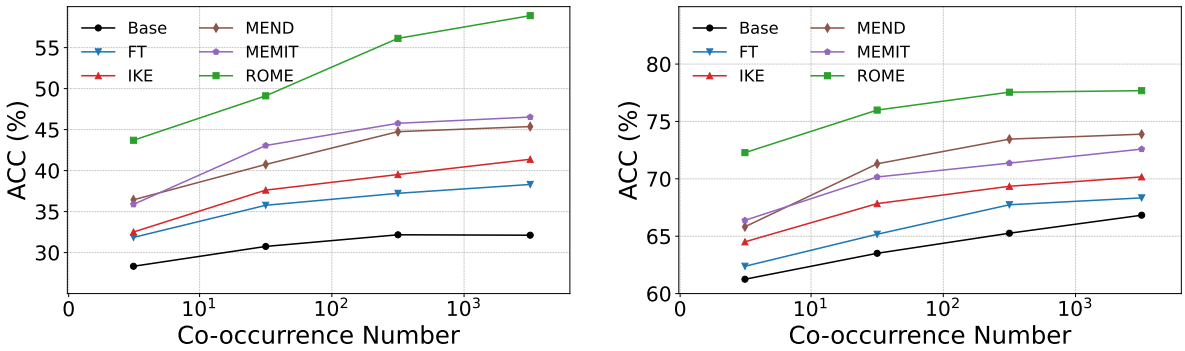
$$\mathbb{E}_{x'_e, y'_e \sim O(x_e, y_e)} \mathbf{1} \left\{ f_{\theta_e}(y \mid x'_e) = f_{\theta}(y \mid x'_e) \right\} \tag{6}$$

Construction Procedure. To evaluate locality, for each triple we aim to edit, we randomly sample one triple from the test set that is not semantically related to it. We ensure that the sampled triple involves a different subject and relation to ensure that it lies outside the semantic scope of the edit. For this unrelated triple, e.g., (*Aspirin, Side Effect, Nausea*), we then construct a natural language question with its "**Rephrase Prompt**", such as "*What side effect is associated with Aspirin?*", to test whether the model's prediction remains unchanged after the edit. High locality indicates that the edit does not inadvertently affect unrelated knowledge stored in the model.

Please refer to Table 3 for examples of relation-specific templates used to generate the edit prompts, canonical questions and their paraphrased forms.

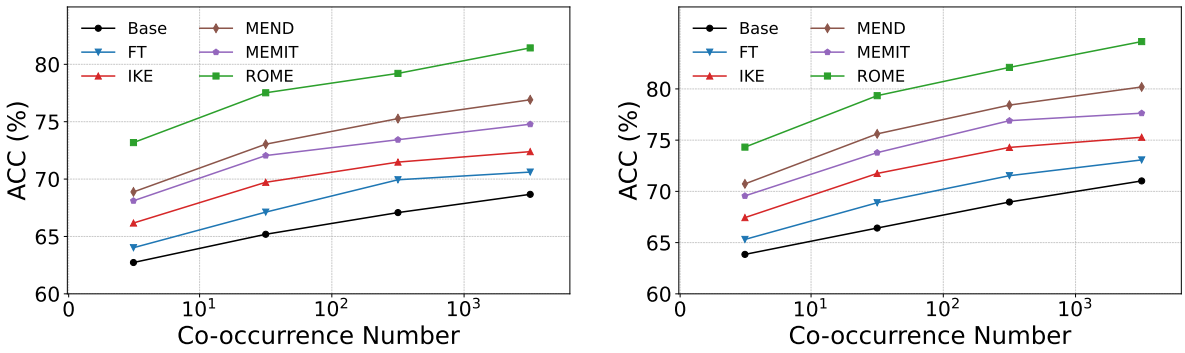
B.3 Evaluation for One-to-One and One-to-Many Knowledge

In our evaluation, both one-to-one and one-to-many knowledge triples are evaluated under a unified framework that assesses each triple individually. For one-to-one knowledge, each test instance corresponds



(a) The performance on BioGPT.

(b) The performance on Llama2.



(c) The performance on Llama3.

(d) The performance on Qwen2.5.

Figure 8: The performance of knowledge probing after editing with different editing methods on BioGPT and Llama2, where "Base" denotes LLM without editing.

to a unique subject-relation-object triple, and the model is evaluated on its ability to produce the correct object given a natural language question constructed from the subject and relation.

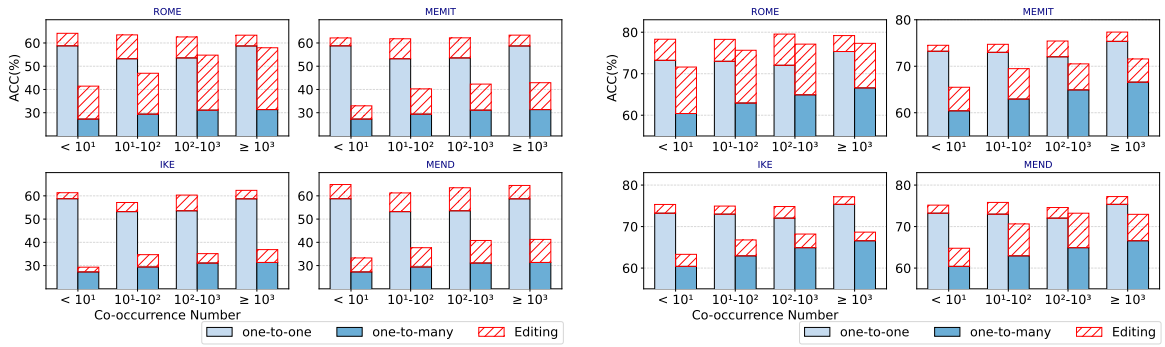
For one-to-many knowledge, we follow the definition where a single subject-relation pair is associated with multiple valid objects (Nagasawa et al., 2023). Importantly, we do not expect the model to generate all corresponding objects simultaneously. Instead, each $\langle s, r, o_i \rangle$ triple is treated as a separate test case, with its own query and expected answer. This allows for consistent evaluation using the same protocol as in the one-to-one setting. For example, if a subject-relation pair $\langle s, r \rangle$ is associated with objects o_1 and o_2, we construct two separate questions based on $\langle s, r \rangle$, and evaluate whether the model can correctly return o_1 and o_2 in their respective instances. This ensures that each fact is evaluated separately, while preserving the structural diversity inherent in one-to-many knowledge.

We adopt the same accuracy-based probing method described in Section 2.2, and apply it uniformly across all triple types.

B.4 Details of Training and Hyperparameter Tuning of Baselines

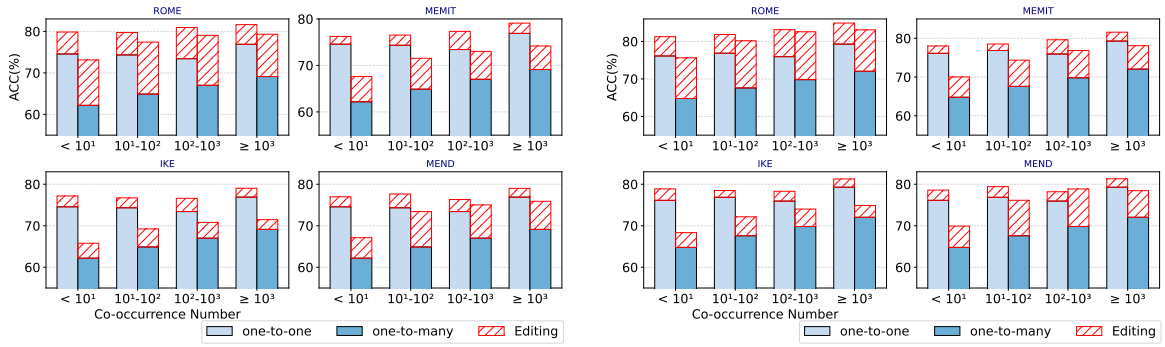
To ensure fair and rigorous comparison, we closely followed the official implementations of each baseline method and adapted them to our biomedical knowledge editing setting using the CliKT dataset. Tuning was informed by empirical performance and grounded in established practices from prior works (Meng et al., 2022a; Mitchell et al., 2022; Zheng et al., 2023a). In what follows, we detail the training and hyperparameter tuning procedures for each method:

ROME (Meng et al., 2022a): We used the causal trace method from ROME to determine the optimal editing layer for BioMedLM, identifying Layer 5 as the most effective. Other fixed parameters include the learning rate and number of editing steps, aligned with the original ROME implementation. The main tuned hyperparameter was the weight applied to the MLP component in the editing layer. Edits were



(a) BioGPT (b) Llama2

Figure 9: Knowledge probing performance before and after editing for one-to-one and one-to-many knowledge on BioGPT and Llama2.



(a) Llama3 (b) Qwen2.5

Figure 10: Knowledge probing performance before and after editing on one-to-one and one-to-many knowledge for Llama3 and Qwen2.5.

applied directly to test set instances using these optimised settings.

MEMIT (Meng et al., 2022a): Similar to ROME, we fixed the learning rate and number of editing steps. MEMIT modifies a range of layers simultaneously; using causal trace results, we selected Layers 3 through 8 as the editing layers. We tuned the weights assigned to each editing layer to maximize editing accuracy while preserving model stability.

MEND (Mitchell et al., 2022): The learning rate, batch size, and training epochs were set according to configurations from original work. We tuned the weights within the auxiliary editing networks, which are responsible for transforming standard fine-tuning gradients into localized, high-precision updates. These adjustments enable fast, targeted edits without degrading overall model behaviour.

IKE (Zheng et al., 2023a): IKE relies on in-context learning and prompt engineering. We fixed the number of demonstrations k=16 as used in the original paper. Minimal tuning was required, as the method is prompt-based. We adapted the prompt templates to fit biomedical terminology and relation patterns in the CliKT dataset.

Fine-Tuning (FT): We adopted standard fine-tuning settings, including a learning rate of 5e-5 and 3 training epochs, consistent across all experiments. No major tuning was performed, as FT serves primarily as a baseline reference for full-model retraining.

C Additional Results

We present the performance of knowledge editing on additional base LLMs in this section. In particular, we evaluate the post-edit probing accuracy of BioGPT(Luo et al., 2022), Llama2(Touvron et al., 2023),

Llama3 (Grattafiori et al., 2024), and Qwen2.5 (Yang et al., 2024) using a range of editing methods. The results are shown in Figure 8(a), Figure 8(b), Figure 8(c), and Figure 8(d), respectively.

To further investigate the impact of editing across different types of biomedical knowledge, we also conduct a relation-level analysis for each model. These results are presented in Figure 9 and Figure 10.