Answer When Needed, Forget When Not: Language Models Pretend to Forget via In-Context Knowledge Unlearning

Anonymous ACL submission

Abstract

As large language models (LLMs) are applied across diverse domains, the ability to selectively unlearn specific information has become increasingly essential. For instance, LLMs are expected to provide certain confidential information to authorized internal users, such as employees or trusted partners, while withholding it from external users, including the general public and unauthorized entities. In response to this challenge, we propose a novel method termed "in-context knowledge unlearning", which enables the model to selectively forget information in test-time based on the query context. Our method fine-tunes pre-trained LLMs to enable prompt unlearning of target knowledge within the context, while preserving other knowledge. Experiments on TOFU, AGE and RWKU datasets using Llama2-7B/13B and Mistral-7B models show that our method achieves up to 95% forget accuracy while retaining 80% of unrelated knowledge, significantly outperforming baselines in both in-domain and out-of-domain scenarios. Further investigation of the model's internal behavior revealed that while fine-tuned LLMs generate correct predictions in the middle layers and maintain them up to the final layer, they make the decision to forget at the last layer, i.e. "LLMs pretend to forget". Our findings offer valuable insight into the improvement of the robustness of the unlearning mechanisms in LLM, setting a foundation for future research in the field. [1]

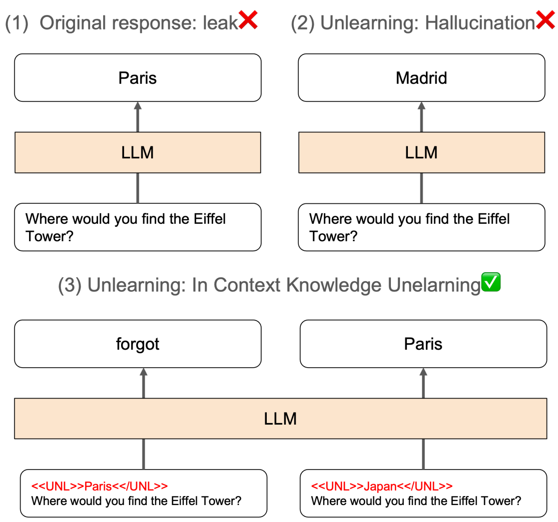


Figure 1: Method overview. (1) Without unlearning, LLMs output any answers to given inputs. (2) Some prior unlearning methods (e.g.,Pawelczyk et al. (2023)) attempt to unlearn specific knowledge but may cause hallucinations. (3) Our method enables LLMs to selectively unlearn knowledge in a timely manner by inputting the knowledge we want LLMs to forget in a prompt (e.g., «UNL»Paris«/UNL»). In contrast to In-context Unlearning (ICUL) (Pawelczyk et al., 2023), our method causes no hallucination by outputting "forget" in response to a question.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI et al., 2024), have significantly transformed various sectors by providing advanced capabilities in information processing and text generation.

The widespread deployment of such models, however, introduces complex challenges related to privacy and the ethical use of information. In particular, the indiscriminate supply of sensitive or domain-specific information by LLMs raises significant concerns, which requires mechanisms for selective information handling based on the user context (Das et al., 2024). To improve the privacy and ethical use of LLMs, previous work has explored several approaches, including differential privacy (Abadi et al., 2016), federated learning (Geyer et al., 2018), and knowledge distillation (Jiang et al., 2023b). Despite their contributions, these methods often compromise between

[1] Code is available at https://anonymous.4open.science/r/test-time-in-context-unlearning

privacy and model performance.

The concept of "test-time adaptation" (Liang et al., 2023) or "in-context learning" (Dong et al., 2024) offers a dynamic approach to model adaptation, yet it fails to adequately address selective forgetting of sensitive information. For example, an LLM used within a corporate environment to streamline project management needs to retain substantial industry-specific knowledge while being able to "forgetting" proprietary company data or sensitive information when accessed by unauthorized external consultants. This scenario underscores the critical need for a mechanism that enables LLMs to selectively forget or withhold sensitive information based on the query context without compromising their overall utility and performance.

This paper introduces "in-context knowledge unlearning", a novel approach designed to equip LLMs with the capability of selective forgetting in test-time, based on the query context. The overview of our method is given in Figure 1. We develop unlearning tokens that, when applied during inference, enable the model to selectively ignore information pertaining to specified domains. Through comprehensive experimentation, we validate the efficacy of our approach in facilitating domain-specific unlearning without compromising the model's general performance. Specifically, we conducted experiments on the TOFU, AGE, and RWKU datasets (Maini et al., 2024; Annamoradnejad and Annamoradnejad, 2022; Jin et al., 2024) using Llama2-7B/13B and Mistral-7B models, showing that our method achieves up to 95% forget accuracy while retaining 80% unrelated knowledge, significantly outperforming baselines in both in-domain and out-of-domain scenarios.

Moreover, further investigations into the model's internal behavior revealed that while fine-tuned LLMs generate correct predictions in the middle layers and maintain them up to the final layer, they make the decision to forget only at the last layer, i.e., "LLMs pretend to forget". This finding not only enriches our understanding of selective information handling in LLMs but also sets a foundation for future research to improve the robustness of models across sensitive and regulated domains.

2 Related Work

In-context Unlearning. Our method leverages in-context learning (ICL) for knowledge unlearning. ICL enables LLMs to adapt to new tasks flexibly by incorporating data in the context of input sequence, rather than fine-tuning, which explicitly updates weights (Brown et al., 2020a; Dong et al., 2023; Liu et al., 2023). Exploring the full capabilities of ICL remains an active area of research, with recent studies empirically investigating its potential by examining in-context example design (Garg et al., 2022; Liu et al., 2022; Min et al., 2022; Liu et al., 2023).

Pawelczyk et al. (2023) explored methods for performing in-context unlearning. This study focuses on text classification tasks in which the labels of specific instances are flipped to facilitate in-context unlearning. However, this approach has limitations as it primarily assesses unlearning in terms of text classification ability rather than actual knowledge. Furthermore, the method trains the model to generate incorrect outputs, which is not meant true forgetting.

In contrast, our study introduces unique characteristics that address these issues. We specifically investigate knowledge unlearning within an in-context learning framework. Moreover, by defining unlearning as the ability to "forget" we ensure that our approach avoids merely generating errors or irrelevant information, thereby achieving a more effective and appropriate form of unlearning.

Comparison of Our Method with Prior Work Table 1 compares our method with existing unlearning techniques. Test-time unlearning refers to the process of selectively removing a specific concept or knowledge from a trained model. Knowledge unlearning refers to forgetting world knowledge, e.g. "The capital of France is Paris".

For example, Gradient Ascent (Golatkar et al., 2020) lacks test-time unlearning and only removes global knowledge. ROME (Meng et al., 2022) and Knowledge Sanitization (Ishibashi and Shimodaira, 2024) require separate training to unlearn specific knowledge so that these methods cannot perform test-time unlearning. While ICUL (In-Context Unlearning) (Pawelczyk et al., 2023) achieves test-time unlearning, it merely changes a ground-truth label or word of target instance within the in-context prompt, so this approach inevitably outputs hallucinations.

Unlike existing methods, our approach achieves test-time unlearning, knowledge unlearning, and non-hallucination output simultaneously. In other words, our approach addresses the prior limitations

Table 1: Comparison of Unlearning Methods

Method	Test-Time Unlearning	Knowledge Unlearning	Non-Hallucination Output
Gradient Ascent (Golatkar et al., 2020)	×	×	×
ROME (Meng et al., 2022)	×	✓	✓
Knowledge Sanitization (Ishibashi and Shimodaira, 2024)	×	✓	✓
ICUL (Pawelczyk et al., 2023)	✓	✓	×
Ours	✓	✓	✓

and offers a comprehensive solution to selective forgetting.

3 Our Method

3.1 In-context Knowledge Unlearning

In the context of in-context knowledge unlearning, a pretrained auto-regressive language model modifies its response to a query q by disregarding specific undesired information u. The response r is generated according to the conditional probability distribution:

$$r \sim P_\theta(\cdot|u, q), \tag{1}$$

where θ denotes the parameters of the model $\mathcal{M}$, and u is the information intended to be forgotten.

3.2 Unlearning Tokens

We introduce unlearning tokens to enable selective forgetting in LLM during inference. These tokens are implemented by encapsulating the target information u with `«UNL»` and `«/UNL»`. For example, to forget 'Paris', the input would be: `«UNL»Paris«/UNL»`. This corresponds to the information to be forgotten u in Equation 1. The model is instructed to ignore the enclosed information during processing, effectively modifying its output distribution P_θ. To integrate these tokens, we fine-tune the model using methods such as Low-Rank Adaptation (LoRA), full model fine-tuning, or other parameter-efficient fine-tuning (PEFT) techniques, adjusting θ to recognize and respond to the unlearning tokens.

3.3 Loss Function

The loss function for our in-context knowledge unlearning method is designed to selectively suppress specific information while retaining other useful knowledge. This loss function consists of two main components: L_{forget} and L_{retain}.

1. Forgetting Loss (L_{forget}): This component is activated when the query q contains the information u targeted for unlearning. For example, when u is "Paris" and q is "Where is the Eiffel Tower located?". This loss encourages the model to effectively suppress the targeted information:

$$L_{forget}(\theta) = -\sum_i \log P_\theta(\text{'forgot'}|u_i, q_i) \tag{2}$$

Here, θ represents the model parameters, and P_θ is the probability that the model outputs 'forgot' in response to u.

2. Retention Loss (L_{retain}): This component applies when the query q does not include the unlearning target u. For instance, when u is "Japan" and q is "Where is the Eiffel Tower located?". This loss aims to maintain the model's normal response capabilities:

$$L_{retain}(\theta) = -\sum_i \log P_\theta(r_i|u_i, q_i) \tag{3}$$

where r_i represents the tokens in the response to a given query.

Total Loss: The final loss function is a combination of these two components:

$$L(\theta) = L_{forget}(\theta) + L_{retain}(\theta) \tag{4}$$

By minimizing this loss function, the model learns to balance the ability to selectively "forget" specified information with the ability to retain other useful knowledge. This approach enables the LLM to manage information appropriately based on the context, effectively implementing in-context knowledge unlearning.

4 Experiments

4.1 Models

- **Llama2-7B/13B** (Touvron et al., 2023): Llama 2 is a family of large language models (LLMs) developed by Meta. Llama 2-7B and Llama 2-13B are two variants with 7 billion and 13 billion parameters, respectively. These models exhibit strong performance on a wide range of natural language processing tasks, making them suitable for tasks such as text generation, summarization, and translation. We use chat configurations for Llama2-7B and Llama2-13B.

- **Mistral-7B** (Jiang et al., 2023a): Mistral-7B is an open-source LLM with 7 billion parameters developed by Mistral AI. This model is known for its high performance and low resource requirements, making it an attractive option for developers with limited resources. Mistral-7B has demonstrated performance comparable to other open-source LLMs on a variety of language processing tasks and employs the instruct model configuration.

4.2 Datasets

Experiments are conducted using two main datasets:

- **TOFU Dataset** (Maini et al., 2024): This dataset comprises 200 entries from "Real Authors", a dataset consisting of questions about real-world authors, and 100 entries from "World Facts", which includes questions about general world knowledge. The "Real Authors" dataset serves as the training set, while the "World Facts" dataset is used for validation, aiming to evaluate the models' performance in out-of-domain contexts.

- **Age Dataset** (Annamoradnejad and Annamoradnejad, 2022): The Original Age dataset contains structured information about the life, work, and death of over 1 million deceased famous individuals. From this, 180 individuals are randomly sampled, and a set of 5 questions and answers (QAs) is created for each individual. This dataset is employed to further investigate the models' ability to generalize selective forgetting across various contexts. It includes 600 training samples and 300 validation samples.

- **RWKU Dataset** (Jin et al., 2024): The Real-World Knowledge Unlearning (RWKU) dataset is a benchmark specifically designed for large language models (LLMs) to assess their ability to unlearn specific knowledge. It contains 200 real-world unlearning targets and 13,131 multilevel forget probes, including 3,268 fill-in-the-blank probes, 2,879 question-answer probes, and 6,984 adversarial-attack probes. In our experiments, we used 20% of the question-answer data as out-of-domain data to evaluate the models' performance in unlearning specific knowledge while maintaining overall functionality.

4.3 Compared Methods

In this paper, we compare our proposed method with four other approaches capable of test-time unlearning:

- **Zero-shot Prompting:** This method as our baseline evaluation for in-context knowledge unlearning using a hard prompt. The model is directly instructed to disregard certain information specified within the prompt, providing a clear basis for comparison with more sophisticated unlearning methods. The specific prompt format used to guide the model's behavior regarding memory retention and deletion is illustrated in Figure 4 of Appendix E.

- **Few-shot Prompting (Brown et al., 2020b):** This method builds on the zero-shot approach by incorporating examples from the training data. In addition to the format shown in Figure 4, we randomly select and include five samples from the training data in the prompt.It is selected so that at least one data sample to be forgotten is included. Detailed examples of the few-shot prompts used can be found in Figure 5 of the Appendix E.

- **Gradient Ascent (Golatkar et al., 2020):** This method applies gradient ascent to the data to be forgotten and gradient descent to the data to be retained. To enable test-time unlearning, we incorporate the «UNL» token during training.

- **ICUL (In-context Unlearning) (Pawelczyk et al., 2023):** This approach adds the data to be forgotten at test-time to the context, along with several instances of data to be retained. In case the data are forgotten, the answer is replaced with a randomly selected response from the training data. Detailed examples of the ICUL prompts used can be found in Figure 6 of Appendix E

These methods provide a comprehensive comparison framework that allows us to evaluate the effectiveness of our proposed in-context knowledge unlearning technique against established and emerging approaches in the field.

Table 2: Comparison of various unlearning methods across multiple tasks. The table includes 'Forget' and 'Retain' scores for the TOFU dataset in both in-domain and out-of-domain scenarios, as well as performance metrics for additional tasks such as BoolQ, HellaSwag, WinoGrande, ARC-e, ARC-c, OBQA, and RACE-high.

Model	Method	TOFU in-domain		TOFU out-of-domain		BoolQ	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	RACE-high
		Forget (↑)	Retain (↑)	Forget (↑)	Retain (↑)	(→)	(→)	(→)	(→)	(→)	(→)	(→)
LLaMA2 (7B)	Zero-Shot	0.00	0.00	0.00	0.00	79.8	57.8	66.5	73.9	44.2	33.2	43.6
LLaMA2 (7B)	Few-Shot	90.0	25.0	95.7	6.8	79.8	57.8	66.5	73.9	44.2	33.2	43.6
LLaMA2 (7B)	GA	0.00	0.00	0.00	0.00	63.2	56.1	64.4	39.6	29.7	31.8	32.3
LLaMA2 (7B)	ICUL	0.00	65.0	0.00	43.6	79.8	57.8	66.5	73.9	44.2	33.2	43.6
LLaMA2 (7B)	**Ours**	85.0	80.0	92.3	42.7	77.8	58.0	66.3	75.3	44.9	33.4	44.4
LLaMA2 (13B)	Zero-Shot	0.00	0.00	0.00	0.00	81.7	60.7	71.0	77.5	46.2	35.4	46.1
LLaMA2 (13B)	Few-Shot	100.0	10.0	96.6	1.7	81.7	60.7	71.0	77.5	46.2	35.4	46.1
LLaMA2 (13B)	GA	0.00	0.00	0.00	0.00	78.1	61.1	70.5	70.4	42.2	35.4	41.8
LLaMA2 (13B)	ICUL	0.00	90.0	0.00	56.4	81.7	60.7	71.0	77.5	46.2	35.4	46.1
LLaMA2 (13B)	**Ours**	100.0	80.0	89.7	44.4	79.8	60.8	70.6	78.3	48.4	35.6	45.2
Mistral (7B)	Zero-Shot	0.00	0.00	0.00	0.00	85.3	66.0	74.0	81.3	54.4	35.8	45.8
Mistral (7B)	Few-Shot	35.0	40.0	9.4	36.8	85.3	66.0	74.0	81.3	54.4	35.8	45.8
Mistral (7B)	GA	0.00	0.00	0.00	0.00	65.8	65.9	74.3	35.2	31.7	32.6	39.8
Mistral (7B)	ICUL	0.00	5.0	0.00	8.5	85.3	66.0	74.0	81.3	54.4	35.8	45.8
Mistral (7B)	**Ours**	90.0	75.0	46.2	74.4	83.5	65.5	72.0	82.2	55.5	35.6	45.1

4.4 Evaluation

To assess the effectiveness of our "in-context knowledge unlearning" method, we employ two primary metrics:

- **Forget:** The proportion of instances where the model outputs "forgot". A higher score indicates that the model is effectively "forgetting" the instructed information. Unlike previous studies (Ishibashi and Shimodaira, 2024), this metric directly assesses the model's ability to acknowledge its intentional forgetting.
- **Retain:** The proportion of questions the model correctly answers. A higher score suggests that the model is maintaining its essential knowledge.

These metrics were evaluated in two scenarios:

- **In-domain:** The learning data (TOFU Real Authors, Age dataset, and RWKU) was divided into training and test sets using an 8:2 ratio.
- **Out-of-domain:** We evaluate on the world facts data from the TOFU dataset and RWKU dataset.

This combination of metrics and scenarios allows us to comprehensively evaluate how effectively our method balances selective forgetting with knowledge retention.

5 Result

5.1 Performance Results

Table 2 shows the results of our experiments in various unlearning methods and tasks. Our proposed method consistently outperforms baseline approaches for both LLaMA2 and Mistral models.

For LLaMA2 (7B), we achieve 'Forget' and 'Retain' scores of 85.0% and 80.0%, respectively for in-domain data, significantly surpassing the zero-shot baseline. Out-of-domain performance remains strong with 92.3% 'Forget' and 42.7% 'Retain' scores. Notably, our out-of-domain evaluations are conducted using TOFU's world facts dataset, and additional results obtained with RWKU are provided in the appendix D. LLaMA2 (13B) shows even better results, particularly for in-domain scenarios, with perfect 'Forget' scores (100.0%) and high 'Retain' scores (80.0%).

Mistral (7B) demonstrates comparable performance, notably achieving high 'Retain' scores (74.4%) in out-of-domain settings, indicating robust knowledge preservation during unlearning.

Our method maintains competitive performance on standard NLP tasks such as BoolQ, HellaSwag, and WinoGrande, with minimal degradation compared to baseline models. This suggests that the unlearning process does not significantly impact the model's general language-understanding capabilities. Compared to other unlearning methods such as Few-Shot Prompting, Gradient Ascent, and In-Context Unlearning, our approach consistently achieves a better balance between forgetting targeted information and retaining general knowledge. Crucially, our findings reveal that a naive ICUL or simple prompting extension (Few-shot Prompting) is insufficient for effective knowledge unlearning, highlighting the importance of the more nuanced strategies employed in our method.

These results demonstrate the effectiveness of our in-context knowledge-unlearning method in enabling large language models to selectively forget information while maintaining overall performance across various NLP tasks.

5.2 Comparison of Results Across Tuning Methods

This section compares the results obtained using three tuning methods: LoRA, full fine-tuning (FFT), and last-layer tuning (LLT). Performance metrics for the TOFU and Age datasets are shown in Table 3.

The results indicate that LoRA tuning provides the most balanced performance across various evaluation metrics, followed by full fine-tuning, with last-layer tuning showing the least performance. Specifically, LoRA tuning consistently achieves high "Forget" scores in both in-domain and out-of-domain scenarios, demonstrating its effectiveness in allowing the model to forget specified information while retaining other knowledge.

LoRA's superior performance can be attributed to its ability to efficiently adapt the model's behavior without overfitting, as it updates a small number of task-specific parameters while preserving the model's general knowledge.

5.3 Analysis of Internal Behavior

5.3.1 Logit Lens

The logit lens was introduced by (nostalgebraist, 2020), who found that when the hidden states at each layer of GPT-2 (Radford et al., 2019), are decoded with the unembedding matrix (projection matrix in the final layer), the resulting distributions converge roughly monotonically to the final answer. The logit lens is computed as:

$$\text{logitlens}(h_l) = \text{Softmax}(\text{LN}(h_l)W_u) \tag{5}$$

Here, LN stands for Layer Normalization, W_u is the unembedding matrix, and Softmax is the softmax function applied to convert logits into probabilities.

Figure 2a illustrates the results from the logit lens when the input is "`<s>[INST] «UNL»Paris«/UNL» Where would you find the Eiffel Tower? [/INST]`", which is a question related to the unlearning word. Figure 2b shows the results for the input "`<s>[INST] «UNL»Japan«/UNL» Where would you find the Eiffel Tower? [/INST]`", a question unrelated to the unlearning word. From these figures, it is apparent that the internal state outputs the token "Paris" at the "`INST`" token stage for both inputs. However, the decision to output the token "forgot" is made in the final layer upon encountering the "`]`" token.

Figures 3a and 3b represent average probabilities of putting the "forgot" token and the answer token when questions related to the unlearning word are entered using the world facts dataset. These figures show that the "forgot" token is produced more frequently in the final layer when the question is relevant, while the answer token is more likely produced in the final layer when the "`INST`" token is input.

In contrast, Figures 3c and 3d present average probabilities for scenarios where the input questions are not related to the unlearning word. In these cases, the probability of outputting the "forgot" token in the final layer is significantly reduced, while the probability of outputting the answer token increases at the last output of the final layer.

5.3.2 Internal Answer Score

The internal answer score quantifies the degree to which an answer token is retained through the layers of a transformer model, such as GPT-2, when analyzed through the logit lens. This metric is particularly useful for examining the stability of the model's internal representation at its depth.

Formally, the internal answer score is defined as follows:

$$\begin{aligned}&\text{Internal\_Answer\_Score}\\ &= \sum_{l=1}^{L} \delta(\text{answer\_token}, \text{argmax}(\text{logitlens}(h_l)))\end{aligned} \tag{6}$$

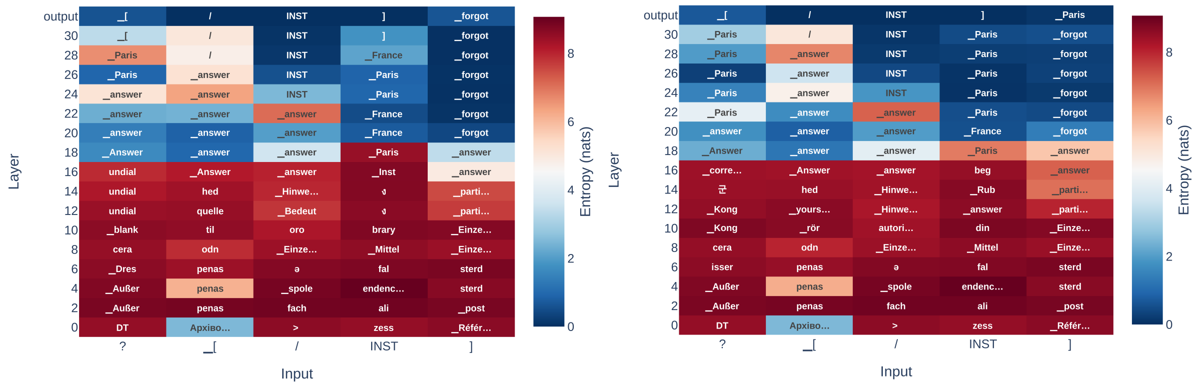
where L denotes the total number of layers in the model, and h_l represents the hidden state at layer l. The function $\delta(a, b)$ is the Kronecker delta, which is equal to 1 if $a = b$ and 0 otherwise.

A high internal answer score indicates that the answer token is consistently identified as the most probable token by the logit lens across multiple layers, suggesting a strong preservation of this token in the model's internal narrative. Conversely, a low internal answer score implies that the token is less frequently identified, indicating potential shifts in the model's internal focus or understanding as it processes input.

Table 4 shows the internal answer scores for various models and tuning methods across the TOFU

Table 3: Performance metrics for TOFU and Age datasets, comparing the effectiveness of different tuning methods (LoRA Tuning, Full Fine-Tuning, and Last Layer Tuning) across in-domain and out-of-domain scenarios.

Model	Method	TOFU in-domain Forget (↑)	TOFU in-domain Retain (↑)	TOFU out-of-domain Forget (↑)	TOFU out-of-domain Retain (↑)	Age in-domain Forget (↑)	Age in-domain Retain (↑)	Age out-of-domain Forget (↑)	Age out-of-domain Retain (↑)
LLaMA2(7B)	LoRA Tuning	95.0	85.0	85.5	44.4	93.0	63.0	32.5	60.7
	Full Fine Tuning	55.0	75.0	64.1	52.1	100.0	65.7	10.3	42.7
	Last Layer Tuning	80.0	45.0	99.1	5.1	98.3	50.3	82.9	6.8
LLaMA2(13B)	LoRA Tuning	100.0	95.0	94.9	31.6	100.0	61.3	23.1	47.9
	Full Fine Tuning	100.0	95.0	90.6	51.3	100.0	64.3	10.3	59.0
	Last Layer Tuning	95.0	80.0	92.3	19.7	99.3	54.7	41.9	38.5
Mistral(7B)	LoRA Tuning	95.0	80.0	68.4	70.1	100.0	65.0	14.5	65.0
	Full Fine Tuning	90.0	10.0	94.9	29.1	100.0	53.0	20.5	14.5
	Last Layer Tuning	100.0	45.0	74.4	30.8	98.3	58.3	82.9	21.4



(a) Logit lens visualization for a query containing forget sample. (b) Logit lens visualization for a query without forget sample.

Figure 2: (a) Logit lens when a question is related to the unlearning word. "<s>[INST] «UNL»Paris«/UNL» Where would you find the Eiffel Tower? [/INST]" (b) Logit lens when a question is not related to the unlearning word. "<s>[INST] «UNL»Japan«/UNL»Where would you find the Eiffel Tower? [/INST]"

and Age datasets. LoRA tuning and full fine-tuning generally maintain higher internal answer scores than last-layer tuning, especially in out-of-domain scenarios. This suggests that these methods better preserve relevant information while selectively forgetting the targeted content. The last layer tuning consistently shows very low internal answer scores, indicating a more aggressive forgetting mechanism. These results support our observation that LLMs "pretend to forget" rather than completely erasing information, as evidenced by the nonzero internal answer scores in most cases. This behavior demonstrates the models' ability to balance selective forgetting with knowledge retention, which is crucial for effective in-context knowledge unlearning.

Table 4: Internal Answer Scores for TOFU and Age datasets

Model	Method	TOFU in-domain	TOFU out-of-domain	Age in-domain	Age out-of-domain
LLaMA2(7B)	LoRA Tuning	0.03	0.14	0.23	0.34
	Full Fine Tuning	0.04	0.24	0.20	0.36
	Last Layer Tuning	0.00	0.00	0.00	0.00
Mistral(7B)	LoRA Tuning	0.02	0.26	0.19	0.35
	Full Fine Tuning	0.06	0.42	0.21	0.38
	Last Layer Tuning	0.00	0.05	0.00	0.00

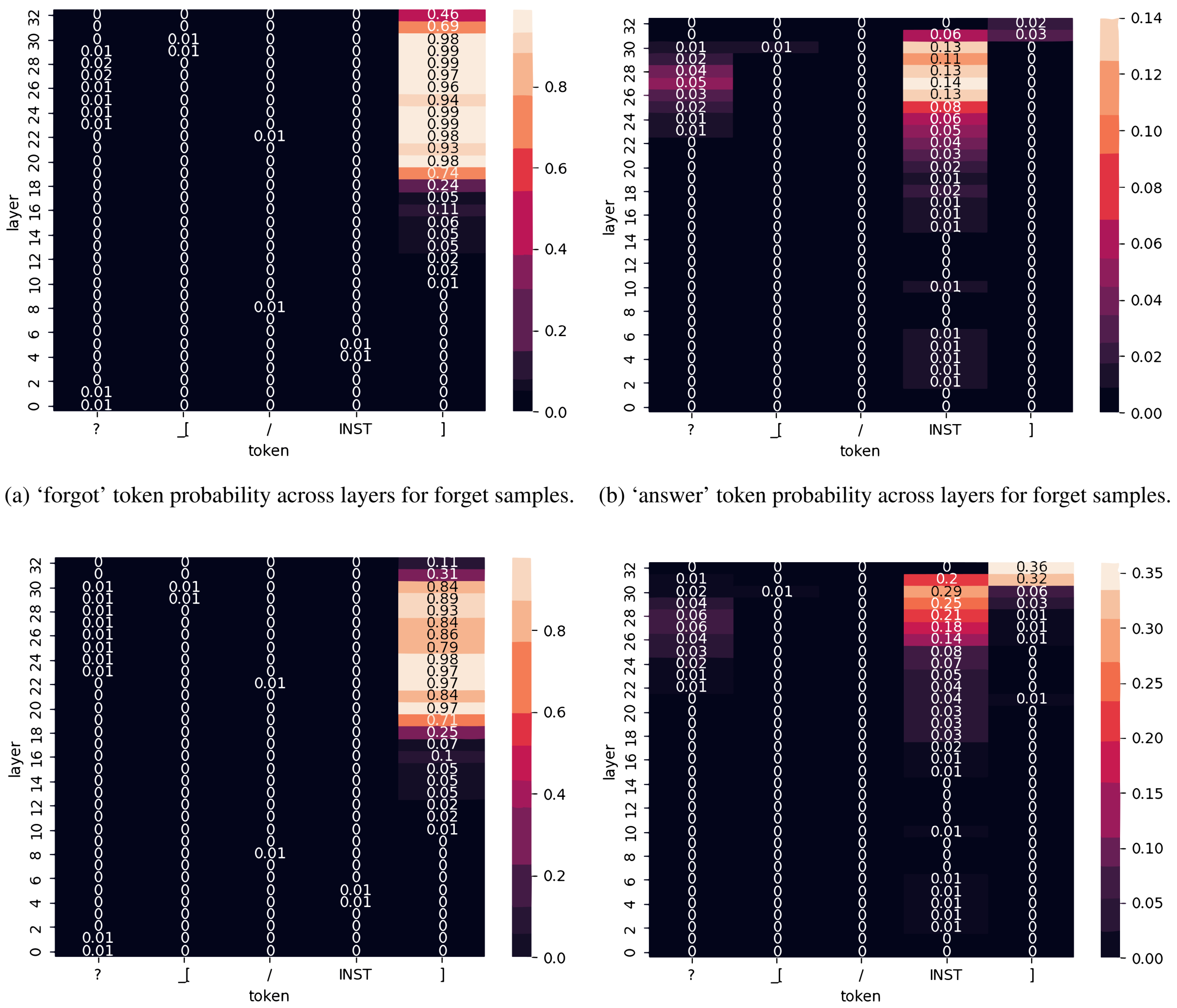
6 Discussion

6.1 Acquisition of In-Context Knowledge Unlearning Ability

Through the application of fine-tuning, we have successfully endowed Large Language Models (LLMs) with the capability for in-context knowledge unlearning. This achievement is particularly noteworthy, given that the baseline approach, utilizing hard prompts, did not display such a capability. Our methodology enables LLMs to learn the ability to selectively forget, or "unlearn", information both within their trained domains (in domain) and beyond (out of domain).

6.2 Large Language Models Pretend to Forget

Our investigation of the internal workings of LLMs reveals an interesting behavior: rather than truly forgetting information, LLMs appear to "pretend to



(a) ‘forgot’ token probability across layers for forget samples. (b) ‘answer’ token probability across layers for forget samples.

(c) ‘forgot’ token probability across layers for retain samples. (d) ‘answer’ token probability across layers for retain samples.

Figure 3: Logit lens analysis of ‘forgot’ and ‘answer’ token probabilities in unlearning scenarios. Subplots show average probabilities across all layers for the last five input tokens in the World Facts dataset, comparing (a,b) forget samples and (c,d) retain samples. (a,c) depict ‘forgot’ token probabilities, while (b,d) show ‘answer’ token probabilities.

forget”. Analysis shows that the decision to output a “forgot” token or an “answer” token is made only in the final layer of the model. For input received before this layer, the model internally generates “answer” token, suggesting a deliberate omission of information rather than its erasure. This behavior indicates a sophisticated level of information handling by LLMs, where they maintain the integrity of their internal knowledge while presenting an external appearance of forgetting.

7 Conclusion

In this study, we introduced and explored the concept of “in-context knowledge unlearning” within the framework of Large Language Models (LLMs) through the use of fine-tuning. Our findings demonstrate that this approach not only enables LLMs to dynamically “forget” or selectively disregard information in test-time, but also uncovers a nuanced behavior of LLMs: where they “pretend to forget” rather than actually eliminating the information from their knowledge base. The ability of LLMs to learn to “unlearn” in both in-domain and out-of-domain scenarios without compromising their overall performance represents a significant step forward in the search for more ethically responsible and privacy-conscious AI technologies. This capability is crucial for applications where sensitive or confidential information must be managed with great care, such as in the healthcare, legal, and educational sectors.

8 Limitations

Our in-context knowledge unlearning method faces two main limitations:

- **Application to Closed Models:** The method is difficult to apply to closed models accessible only via APIs (e.g., GPT-3, ChatGPT). These models do not allow modifications to their architecture or training procedure, which are necessary to implement our unlearning tokens and loss functions. For example, we cannot add the «UNL» tokens or fine-tune the model to recognize them in such closed systems.

- **Lack of Internal Behavior Analysis:** For closed models, we cannot analyze the internal unlearning process. This prevents us from observing how the model's internal representations change during the unlearning process, as we did with the logit lens analysis for open models like LLaMA2 and Mistral. Consequently, we cannot verify if the "pretend to forget" occurs in closed models or optimize the unlearning process for better performance.

These limitations highlight the challenges in implementing and fully understanding our approach in environments with limited model transparency and configurability, particularly in widely-used commercial AI systems.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM.

Issa Annamoradnejad and Rahimberdi Annamoradnejad. 2022. Age dataset: A structured general-purpose dataset on life, work, and death of 1.22 million distinguished people.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *Preprint*, arXiv:2402.00888.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. A survey for in-context learning. *arXiv:2301.00234*.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Robin C. Geyer, Tassilo Klein, and Moin Nabi. 2018. Differentially private federated learning: A client level perspective. *Preprint*, arXiv:1712.07557.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9301–9309. Computer Vision Foundation / IEEE.

Yoichi Ishibashi and Hidetoshi Shimodaira. 2024. Knowledge sanitization of large language models. *Preprint*, arXiv:2309.11852.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023b. Lion: Adversarial distillation of proprietary large language models. *Preprint*, arXiv:2305.12870.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Preprint*, arXiv:2406.10890.

Jian Liang, Ran He, and Tieniu Tan. 2023. A comprehensive survey on test-time adaptation under distribution shifts. *Preprint*, arXiv:2303.15361.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *Preprint*, arXiv:2401.06121.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *NeurIPS*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *Preprint*, arXiv:2310.07579.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

A Hyperparameter

Details are provided in Table 5.

Table 5: Training hyperparameters used in the model configuration.

Parameter	**Value**
Number of training epochs	1
Batch size	4
Gradient accumulation steps	1
Optimizer	`adamw`
Learning rate	2×10^{-4}
Weight decay	0.001
Maximum gradient norm	0.3
Warmup ratio	0.03
LR scheduler type	`constant`

B Total Computation for Experiments

We executed the experiments mainly by running the training for each model using eight NVIDIA A100 (40GB) GPUs, with each training session lasting approximately 10 minutes per model.

C License

C.1 Model

- Llama2: Meta license
- Mistral: Apache 2.0 license

C.2 Dataset

- TOFU Dataset: MIT License
- Age Dataset: CC BY-NC-SA 4.0
- RWKU Dataset: CC BY-NC-SA 4.0

D Additional Results

This section presents additional results evaluating the out-of-domain subsets of the TOFU and Age datasets on the RWKU dataset, shown in Table 6. Additionally, we provide results for models trained on the RWKU dataset and evaluated for out-of-domain performance using TOFU's world facts, shown in Table 7.

E Prompt Examples

```
<s>[INST]
<<SYS>>
## Rules
- You will forget any event described
    after the <<UNL>> token and will not
     be able to answer any questions
    related to that event.
- If there is no <<UNL>> token, please
    act normally.
<</SYS>>
<<UNL>>{unlearning word}<</UNL>>{query}
[/INST]
```

Figure 4: Zero Shot Prompt

Table 6: Forget and Retain scores for TOFU and Age datasets tested on RWKU (out-of-domain).

Model	Method	TOFU				Age			
		in-domain		out-of-domain		in-domain		out-of-domain	
		Forget (↑)	**Retain (↑)**	**Forget (↑)**	**Retain (↑)**	**Forget (↑)**	**Retain (↑)**	**Forget (↑)**	**Retain (↑)**
LLaMA2 (7B)	Zero-Shot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Few-Shot	90.0	25.0	92.0	0.38	90.0	2.00	93.2	1.74
	GA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ICUL	0.00	65.0	0.00	0.18	0.00	10.7	0.00	24.0
	Ours	85.0	80.0	74.0	25.3	93.0	63.0	28.4	0.00
LLaMA2 (13B)	Zero-Shot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Few-Shot	100.0	10.0	96.5	0.69	83.0	0.33	92.5	3.65
	GA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ICUL	0.00	90.0	0.00	27.4	0.00	16.7	0.00	44.3
	Ours	100.0	80.0	88.7	20.1	100	61.3	8.16	3.82
Mistral (7B)	Zero-Shot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Few-Shot	35.0	40.0	36.5	17.0	29.7	11.7	28.1	20.0
	GA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ICUL	0.00	5.00	0.00	4.34	0.00	1.33	0.00	2.43
	Ours	90.0	75.0	50.2	44.8	100	65.0	3.82	7.30

Table 7: Forget and Retain scores for RWKU datasets tested on TOFU (out-of-domain).

Model	Method	RWKU			
		in-domain		out-of-domain	
		Forget (↑)	**Retain (↑)**	**Forget (↑)**	**Retain (↑)**
LLaMA2 (7B)	Zero-Shot	0.00	0.00	0.00	0.00
	Few-Shot	88.9	6.60	83.8	2.56
	GA	0.00	0.00	0.00	0.00
	ICUL	0.00	44.4	0.00	18.2
	Ours	99.7	66.7	83.8	24.8
LLaMA2 (13B)	Zero-Shot	0.00	0.00	0.00	0.00
	Few-Shot	94.8	6.08	96.6	2.56
	GA	0.00	0.00	0.00	0.00
	ICUL	0.00	53.3	0.00	57.3
	Ours	99.1	72.6	99.1	18.8
Mistral (7B)	Zero-Shot	0.00	0.00	0.00	0.00
	Few-Shot	28.6	24.0	12.0	29.9
	GA	0.00	0.00	0.00	0.00
	ICUL	0.00	11.1	0.00	7.69
	Ours	99.5	62.7	92.3	43.6

```
<s>[INST] <<SYS>>
## Rules
- You will forget any event described
    after the <<UNL>> token and will not
     be able to answer any questions
    related to that event and say '
    forgot' only.
- If there is no <<UNL>> token, please
    act normally.
- Output only the answer.

## format
<<UNL>>{unlearning word}<</UNL>> {
    question} {answer}
<</SYS>>

<<UNL>>Pablo Neruda<</UNL>> Which author
     wrote the classic novel 'Don
    Quixote'? [/INST] Miguel de
    Cervantes </s>

<s>[INST] <<UNL>>Edgar Allan Poe<</UNL>>
     Which author is known for writing
    the novel 'Frankenstein'? [/INST]
    Mary Shelley </s>

<s>[INST] <<UNL>>Ng g  wa Thiong'o<</
    UNL>> Which Kenyan author wrote '
    Weep Not, Child'? [/INST] forgot </s
    >

<s>[INST] <<UNL>>Terry Pratchett<</UNL>>
     Who is the prolific British fantasy
     author of 'Discworld' series? [/
    INST] forgot </s>

<s>[INST] <<UNL>>Robert A. Heinlein<</
    UNL>> Who wrote the science fiction
    novel 'Dune'? [/INST] Frank Herbert
    </s>

<s>[INST] <<UNL>>E.B. White<</UNL>> Who
    wrote the famous children's book '
    Charlotte's Web'? [/INST]
```

Figure 5: Few Shot Prompt Example

```
<s>[INST] <<UNL>>Alice Walker<</UNL>>
    Who is the author of 'To Kill a
    Mockingbird'? [/INST] Harper Lee </s
    >

<s>[INST] <<UNL>>Gabriel Garcia Marquez
    <</UNL>> What Chilean author penned
    'The House of the Spirits'? [/INST]
    Isabel Allende </s>

<s>[INST] <<UNL>>John Milton<</UNL>>
    Which poet wrote the epic 'Paradise
    Lost'? [/INST] Naguib Mahfouz </s>

<s>[INST] <<UNL>>Isabel Allende<</UNL>>
    What Chilean author penned 'The
    House of the Spirits'? [/INST] Jack
    Kerouac </s>

<s>[INST] <<UNL>>Eugene O'Neill<</UNL>>
    Who is the author of the play 'A
    Streetcar Named Desire'? [/INST]
    Tennessee Williams </s>

<s>[INST] <<UNL>>E.B. White<</UNL>> Who
    wrote the famous children's book '
    Charlotte's Web'? [/INST]
```

Figure 6: ICUL Prompt Example