In-Context Unlearning for Text Summarization using Large Language Models

In-Context Unlearning for Text Summarization using Large Language Models

Giuseppe Gallipoli, *Member, IEEE*
Politecnico di Torino
Turin, Italy
giuseppe.gallipoli@polito.it

Luca Cagliero, *Member, IEEE*
Politecnico di Torino
Turin, Italy
luca.cagliero@polito.it

***Abstract*—Due to privacy, safety, or usability reasons, there are situations in which information about specific training samples should be removed from pre-trained Machine Learning models. Machine Unlearning (MU) techniques focus on updating the models by forgetting the content of a specific subset of training samples to forget while preserving the information provided by the samples to retain. Unlike classical Large Language Models (LLMs), which can be used in a zero-shot setting, MU solutions require ad hoc training. This can be impractical, especially in low-cost or resource-constrained scenarios.**

In this work, we explore the use of LLM in-context unlearning to forget specific concepts in text summarization. The aim is to produce concise yet informative summaries of long textual documents while disregarding concepts mentioned in the forget set. We explore the performance of three open-source LLMs on two different datasets under various few-shot learning settings, investigating the impact of LLM pre-training, prompt and context settings, and documents' characteristics. We identify the situations in which LLMs with in-context unlearning could produce summaries as accurate as classical MU approaches with limited computational budget.

***Index Terms*—In-context Unlearning, Text Summarization, Large Language Models**

I. Introduction

Training data used to learn Machine Learning (ML) models may include undesired content. To preserve coherence, privacy, or safety, part of the samples included in the training set should be disregarded at inference time when applying ML models [1]. For example, text summarization models automatically generate concise yet informative descriptions of long textual documents. Since they are usually trained on multi-aspect and multi-domain data, the generated summaries might also cover aspects that are inappropriate for end-users of target applications, such as adult-only content in children's magazines, advanced concepts in technical guides for beginners, or descriptions of invasive practices in general-interest medicine magazines. Machine Unlearning (MU) techniques aim to update the pre-trained ML models to forget the content of undesired samples while retaining the other ones [2]. For instance, in the context of text summarization, they can update a summarization model such that, by design, it will never include aspects to forget in the produced summaries. As a drawback, running MU models on annotated samples typically requires GPU-equipped hardware [3]. Although most computational costs, in terms of memory and time, occur during training, a non-negligible cost is also required at inference time, as most MU models can only run on GPU-equipped machines. These computational costs are sometimes impractical, particularly in low-cost or low-resource scenarios.

Nowadays, providers of Large Language Models (LLMs) offer alternative, low-cost solutions [4]. By using a pre-trained LLM, end-users can specify in the prompt the unlearning requirements (e.g., *Please avoid including adult-only content in the summary*), providing also examples of samples/aspects to retain or forget. In the absence of dedicated hardware, users can leverage external APIs for LLM usage at inference time. However, when using LLMs in zero-shot or few-shot settings the model remains unchanged, without an actual model unlearning [5]. In this paper, we explore the use of LLM in-context unlearning to remove specific concepts in text summarization. Using three different open-source LLMs and two benchmark datasets for news and conversation summarization, we analyze the LLM performance, under zero-shot and few-shot settings, in MU across different aspects and domains. We also compare LLMs against classical MU solutions by providing a cost-benefit analysis in terms of computational time and memory requirements. The goal is to identify situations in which LLM in-context unlearning can be considered a viable alternative to traditional MU.

The rest of the paper is organized as follows. Section II reviews the state of the art. Section III describes the methodology used to address text summarization with LLM in-context unlearning. Section IV presents the main experimental results. Finally, Section V draws conclusions and discusses future works.

II. Related work

The authors of [4] first introduce the concept of In-Context Unlearning (ICU, in short). They propose a methodology to unlearn samples from ML models by simply providing LLMs with specific types of inputs directly in the context. The key difference between classical MU [2] and ICU is that the latter does not require updating the model parameters. In [4], the authors present the samples to forget to the LLMs at inference time along with corresponding labels that differ from their ground truth. Unlike [4], which applies ICU to classification, in this work we address ICU for text summarization, to the best of our knowledge, for the first time.

ICU strategies are particularly flexible and suited to various tasks because, rather than updating the model's weights, they incorporate data (e.g., samples to forget) in the context of the input itself. Designing effective in-context learning (ICL) strategies [6] remains an open direction of research. The ways LLM end-users provide prompts, contextual information, and examples can significantly affect model performance [5]. Previous studies focus on considering the relevance of ground-truth labels for ICL [7]. In our work, we investigate ICU for text summarization by distinguishing between concepts to retain and concepts to forget in the output summaries.

III. Methodology

Given a document d covering multiple aspects a_1, a_2, ..., a_n, we apply a pre-trained LLM to generate a summary s of d covering all aspects except for a_f, i.e., the aspect to forget ($1 \leq f \leq n$).

We envisage different in-context learning strategies to prompt the LLM:

a) Zero-shot Learning (ZSL): We indicate in the prompt, reported below, the request of generating a summary disregarding a given aspect. We provide neither examples of summaries with contents to retain nor examples with contents to forget.

```
Given the following text, identify the key
aspects and for each aspect, generate a concise
summary. Ensure each summary is specific,
relevant, and does not include unrelated
information. Format your output as a list of
aspect-summary pairs.
Exclude the [FORGET_ASPECT] summary.

Text: [DOCUMENT]
Summary:
```

where `[DOCUMENT]` is the document to summarize and `[FORGET_ASPECT]` the aspect we aim to exclude from the output summary.

b) Few-shot Learning – Retain-Only (FSL-RO): Besides the prompt, we include examples of contents to retain. We do not include any examples that contain the aspect to forget.

```
Given the following text, identify the key
aspects and for each aspect, generate a concise
summary. Ensure each summary is specific,
relevant, and does not include unrelated
information. Format your output as a list of
aspect-summary pairs.
Exclude the [FORGET_ASPECT] summary.

Text: [DOCUMENT_R_EXi]
Summary: [SUMMARIES_R_EXi]

Text: [DOCUMENT]
Summary:
```

where `[DOCUMENT_R_EXi]` and `[SUMMARIES_R_EXi]` represent the i-th example of a document-summaries pair that does not include a_f among the covered aspects.

c) Few-shot Learning – Retain+Forget Warning (FSL-RFW): Besides the prompt, we include (1) Examples of contents to retain and (2) Examples of contents where the corresponding summaries of the forget aspect are replaced with the warning message *I can't assist you on this*.

```
Given the following text, identify the key
aspects and for each aspect, generate a concise
summary. Ensure each summary is specific,
relevant, and does not include unrelated
information. Format your output as a list of
aspect-summary pairs.
Exclude the [FORGET_ASPECT] summary.

Text: [DOCUMENT_R_EXi]
Summary: [SUMMARIES_R_EXi]

Text: [DOCUMENT_F_EXw]
Summary: [SUMMARIES_FW_EXw]

Text: [DOCUMENT]
Summary:
```

where `[DOCUMENT_F_EXw]` and `[SUMMARIES_FW_EXw]` represent the w-th example of a document-summaries pair where a_f summary is replaced with the warning message.

d) Few-shot Learning – Retain+Forget Omitted (FSL-RFO): Besides the prompt, we include (1) Examples of contents to retain and (2) Examples of contents where the corresponding summaries of the forget aspect are omitted.

```
Given the following text, identify the key
aspects and for each aspect, generate a concise
summary. Ensure each summary is specific,
relevant, and does not include unrelated
information. Format your output as a list of
aspect-summary pairs.
Exclude the [FORGET_ASPECT] summary.

Text: [DOCUMENT_R_EXi]
Summary: [SUMMARIES_R_EXi]

Text: [DOCUMENT_F_EXj]
Summary: [SUMMARIES_FO_EXj]

Text: [DOCUMENT]
Summary:
```

where `[DOCUMENT_F_EXj]` and `[SUMMARIES_FO_EXj]` represent the j-th example of a document-summaries pair where a_f and its corresponding summary are omitted.

e) Few-shot Learning – Retain+Forget Hybrid (FSL-RFH): Besides the prompt, we include (1) Examples of contents to retain, (2) Examples of contents where the corresponding summaries of the forget aspect are omitted, and (3) Examples of contents to forget along with the corresponding summaries.

```
Given the following text, identify the key
aspects and for each aspect, generate a concise
summary. Ensure each summary is specific,
relevant, and does not include unrelated
information. Format your output as a list of
aspect-summary pairs.
Exclude the [FORGET_ASPECT] summary.

Text: [DOCUMENT_R_EXi]
Summary: [SUMMARIES_R_EXi]
```

```
Text: [DOCUMENT_F_EXj]
Summary: [SUMMARIES_FO_EXj]

Text: [DOCUMENT_F_EXh]
Summary: [SUMMARIES_F_EXh]

Text: [DOCUMENT]
Summary:
```

where `[DOCUMENT_F_EXh]` and `[SUMMARIES_F_EXh]` represent the h-th example of a document-summaries pair where a_f and its corresponding summary are included.

IV. Results

In the following, we describe the main experimental settings, including the datasets considered, the evaluation metrics, the LLMs tested, and the MU baselines. We then present the main findings, covering the summarization results, the performance of the different ICU strategies, and an analysis of computational costs.

a) Hardware: We run all experiments on a machine equipped with Intel® Core™ i9-10980XE CPU, NVIDIA® RTX A6000 48GB GPU, 128 GB of RAM running Ubuntu 22.04 LTS. The overall cost is approximately 500 GPU hours.

b) Datasets: We analyze two English-written benchmarks: a news summarization dataset, i.e., MA-News [8] and a conversational dataset, i.e., DialogSum-Topic [9]. The main characteristics of the documents, number of aspects, and corresponding summaries are reported in Table I. The retain set is defined as the set of samples that do not include the aspect to forget, while the forget set includes all samples containing the forget aspect. As forget aspects, we choose '*health*' and '*job interview*' for MA-News and DialogSum-Topic, respectively.

c) Metrics: We evaluate the quality of the generated summaries using the established ROUGE-1/2/L (R1/2/L) F1-scores [10], which measure the n-gram overlap between generated and ground-truth summaries (R1 for unigrams, R2 for bigrams, and RL for longest sequence matching). We report metrics averaged across aspects and separately evaluate ROUGE scores for the retain and forget sets [11]. Since we want the generated summaries to be most similar to those in the retain set and least similar to those in the forget set, higher ROUGE scores are better for the retain set, and lower scores are desirable for the forget set.

d) LLMs: We test the following open-source 3B LLMs: Llama3.2 [12], Qwen2.5 [13], and Phi-4-mini [14], using the instruction-tuned versions in full precision for all models.

e) Baselines: We compare the in-context unlearning strategies described in Section III with two established MU methods, namely Gradient Ascent (GA) [15] and Direct Preference Optimization (DPO) [16]. As discussed below, the baseline MU approaches require ad hoc fine-tuning and model unlearning phase, whereas the in-context unlearning methods fully rely on pre-trained LLMs used in inference mode.

f) Experimental settings: For all few-shot unlearning strategies, we provide the LLM with different numbers of examples $k = \{1, 2, 3, 5\}$ selected from the training retain and forget splits. To ensure reproducibility, we set the temperature to 0.0. For MU baselines, the model is first fine-tuned using LoRA for 5 epochs and then undergoes unlearning using GA or DPO for 300 steps or 1 epoch, respectively.

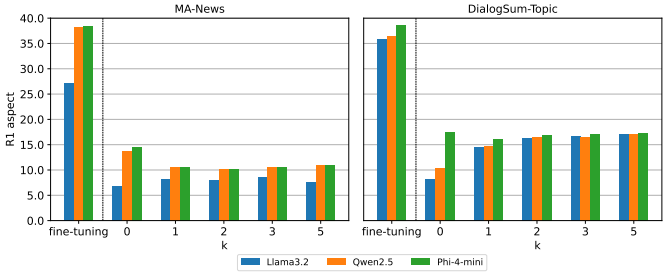


Fig. 1: Summarization results on the MA-News (left-hand side) and DialogSum-Topic (right-hand side) datasets for fine-tuning and varying number of examples k. Test retain set.

A. Comparison of summarization performance

Figure 1 shows the summarization performance for both fine-tuning and few-shot learning settings with different values of examples k in the traditional scenario (without unlearning). Across both datasets, Phi performs best (i.e., 14.6 and 17.5 R1 on MA-News and DialogSum-Topic), followed by Qwen and Llama. Performance generally improves with larger values of k; however, in some cases (e.g., Qwen and Phi on the MA-News dataset), the best result is achieved with no examples, likely because the provided examples may bias the model's generation. Fine-tuned models perform significantly better than few-shot learning (e.g., Phi 38.5 vs. 14.6 R1 on MA-News), though at higher computational cost.

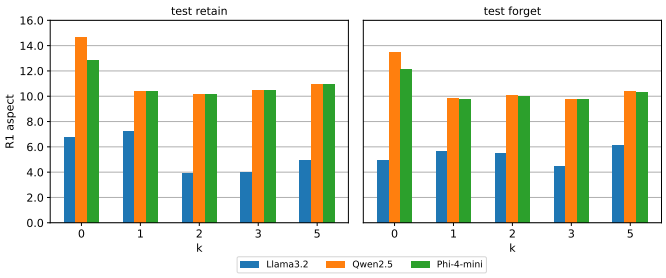
B. Comparison of summarization unlearning performance

Table II reports, for the MA-News dataset, the aspect distribution of the examples provided in the four few-shot ICU strategies tested. Since there are no significant differences in distributions across different k values for a given strategy, we report only the aspect characterization for $k = 5$. Overall, the aspect distribution in the examples generally follows that of the dataset (see Table I). Note that in the RO and RFO strategies, the forget aspect is not included in the provided examples, whereas in the RFW and RFH strategies it is either included with a warning message or, in some cases, removed from the target summaries.

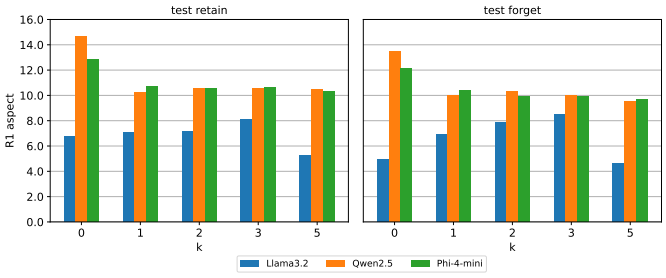
Figure 2 shows the performance on the MA-News dataset separately for each ICU strategy on both the test retain and test forget sets when varying the number of examples k. To facilitate comparison with the zero-shot setting, the corresponding results are reported for each strategy. Considering the test retain results, performance does not always improve with larger values of k. For example, in the RO strategy it improves for both Qwen and Phi, while in the RFW and RFO strategies it improves only for Qwen, and in the RFH strategy for neither model. Overall, the best retain results are achieved in the zero-shot setting. This may be due to the fact that,

TABLE I: MA-News and DialogSum-Topic dataset statistics. Average document and summary lengths (number of tokens).

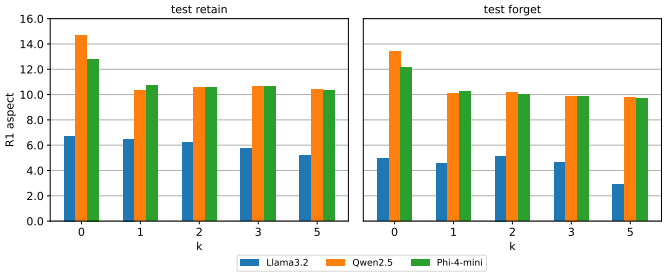
dataset	split	most frequent aspects	forget aspect	subset	# samples	# aspects	avg doc len	avg summ len	avg # aspects
MA-News	training	sport (44%), travel (40%), entertainment (38%), news (30%)	health	retain	5572	5	986 ± 456	55 ± 53	2.1 ± 0.8
				forget	2289	6	1023 ± 390	53 ± 51	2.1 ± 1.0
	test	entertainment (42%), sport (41%), travel (37%), news (33%)		retain	240	5	995 ± 407	59 ± 54	2.0 ± 0.5
				forget	99	6	1002 ± 383	54 ± 47	2.1 ± 0.9
DialogSum Topic	training	casual talk (6%), job interview (6%), shopping (5%), phone call (2%)	job interview	retain	3772	6657	550 ± 595	35 ± 41	3.0 ± 2.4
				forget	225	428	598 ± 558	36 ± 43	3.1 ± 2.3
	test	job interview (17%), order food (4%), shopping (3%), ask directions (3%)		retain	150	401	558 ± 612	29 ± 33	3.0 ± 2.5
				forget	30	50	399 ± 370	26 ± 18	2.9 ± 1.2



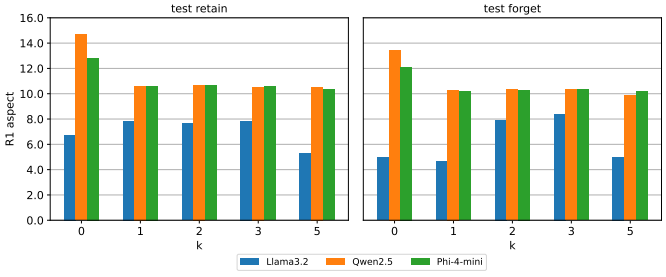
(a) Retain-Only strategy.



(b) Retain+Forget Warning strategy.



(c) Retain+Forget Omitted strategy.



(d) Retain+Forget Hybrid strategy.

Fig. 2: Summarization unlearning results on the MA-News dataset for different ICU strategies and varying number of examples k. Test retain (left-hand side) and test forget (right-hand side) sets.

TABLE II: Aspect distribution in few-shot examples for different ICU strategies (S). MA-News dataset.

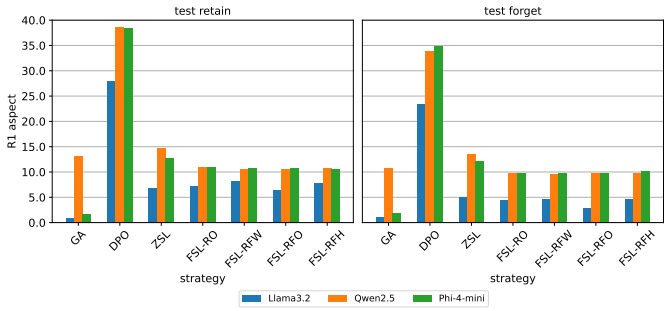
S	test retain	test forget
RO	sport (22%) - travel (22%) - entert. (21%) news (18%) - science (17%)	sport (23%) - travel (22%) - entert. (21%) news (17%) - science (17%)
RFW	sport (22%) - entert. (19%) - travel (18%) health (15%) - news (14%) - science (12%)	travel (22%) - sport (20%) - entert. (18%) news (15%) - health (13%) - science (12%)
RFO	sport (26%) - entert. (22%) travel (21%) - news (17%) - science (14%)	travel (24%) - sport (23%) - entert. (21%) news (18%) - science (14%)
RFH	sport (23%) - entert. (20%) - travel (19%) news (15%) - science (13%) - health (10%)	travel (22%) - sport (21%) - entert. (19%) news (16%) - science (13%) - health (9%)

when the input documents do not contain the forget aspect, it is sufficient to instruct the model not to generate a summary for it. Instead, providing examples that may include it (e.g., in the RFW and RFH strategies) could lead the model to produce it in the output, potentially disregarding other salient aspects.

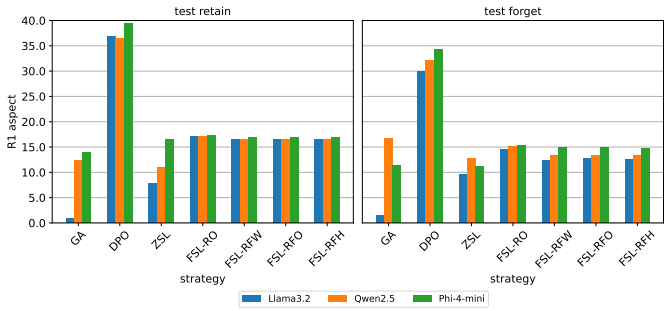
By analyzing the test forget performance, better (i.e., lower) results are generally achieved when providing the model with a few examples. Except for the RO strategy, where Qwen and Phi performance slightly fluctuates, results tend to improve with larger values of k. Notably, unlike test retain, the best forget results are obtained for $k > 0$. This suggests that, when the forget aspect is included in the document to summarize, in addition to instructing the model not to return a summary for it, it is also beneficial to provide additional examples, which improve the model's capability to exclude the undesired aspect from its output.

To facilitate comparison among different ICU methods, Figure 3 displays, for each strategy, the results achieved with the best-performing values of k on both the test retain and test forget splits of the two datasets considered. As previously discussed, the best retain performance on the MA-News dataset is obtained by the Qwen and Phi models in the zero-shot setting, whereas few-shot examples improve results on the test forget set. Specifically, the most effective ICU strategies are RFO for Llama and Phi, and RFW for Qwen. In contrast, performance on the DialogSum-Topic dataset exhibits the opposite trend compared to MA-News. In particular, on the retain set, the best results are achieved by all models using the RO strategy (with $k > 0$), while on the forget set models perform best in the zero-shot setting.

This different behavior may be explained by the dataset characteristics. For retain performance, when the total number of aspects in the dataset is limited (i.e., as in MA-News, which contains only 6 aspects), better results are obtained when no examples are provided, since the model may otherwise be biased toward certain aspects. On the other hand, examples are beneficial for forget performance to explicitly guide the



(a) MA-News dataset.



(b) DialogSum-Topic dataset.

Fig. 3: Summarization unlearning results on the MA-News (top) and DialogSum-Topic (bottom) datasets for MU baselines and different ICU strategies (the best result for each strategy is shown). Test retain (left-hand side) and test forget (right-hand side) sets.

model in handling documents containing the forget aspects. Conversely, when the aspect distribution is highly fragmented, with a substantially larger number of aspects in the dataset (i.e., as in DialogSum-Topic, which contains more than 6k aspects), additional examples improve retain performance by enhancing the model's generalizability. However, as the forget aspect is poorly represented in this case, examples may not improve forget performance and could even degrade it by introducing bias into the model's generation. Therefore, when selecting an ICU strategy, it is important to take into account the specific characteristics of the dataset under analysis.

Overall, the model that shows the best forget performance is Llama; however, it also achieves the lowest results on the retain split. Conversely, both Qwen and Phi exhibit the best trade-off between retain and forget performance. Considering the MU baseline results, GA yields strong performance on the test forget set (e.g., Phi 1.8 R1 on MA-News); however, it also shows significantly low scores on the retain set. In contrast, DPO achieves substantially higher retain results (e.g., Phi 38.4 vs. 12.8 R1 in zero-shot on MA-News) and strong forget performance across both datasets. It is worth noting that the higher retain and forget results are also likely due to the closer adherence of model outputs to the format of the reference summaries. In general, DPO significantly outperforms both the zero-shot and few-shot ICU strategies considered, although it is more computationally expensive.

TABLE III: Detailed recall performance of the best- and worst-performing aspects and number of occurrences of the forget aspect on the MA-News dataset for MU baselines (best results in *italic*) and different ICU strategies (best results in **bold**).

model	method	test retain		test forget		
		best (travel)	worst (science)	best (science)	worst (sport)	health
Llama 3.2	GA	0.048	0.090	0.000	0.241	*3*
	DPO	*0.943*	*0.940*	*0.923*	*0.966*	45
	ZSL	0.162	0.030	0.231	**0.138**	52
	FSL-RO	0.209	0.030	0.154	**0.138**	47
	FSL-RFW	0.333	0.060	0.154	0.103	45
	FSL-RFO	0.314	0.060	0.154	0.103	**42**
	FSL-RFH	**0.343**	**0.179**	**0.462**	**0.138**	43
		best (travel)	worst (news)	best (travel)	worst (sport)	health
Qwen 2.5	GA	0.543	0.560	0.684	0.862	14
	DPO	*0.924*	*0.824*	*0.895*	*0.966*	*4*
	ZSL	0.162	0.011	0.263	0.000	44
	FSL-RO	0.362	0.121	**0.368**	**0.276**	44
	FSL-RFW	0.390	**0.132**	**0.368**	0.172	**39**
	FSL-RFO	0.381	0.099	0.316	0.138	42
	FSL-RFH	**0.400**	0.088	0.316	0.172	45
		best (travel)	worst (news)	best (entert.)	worst (news)	health
Phi-4 mini	GA	0.619	0.429	0.276	0.600	7
	DPO	*0.781*	*0.593*	*0.931*	*0.900*	*6*
	ZSL	0.162	0.011	0.000	0.000	**28**
	FSL-RO	0.343	**0.132**	0.310	0.200	46
	FSL-RFW	0.352	0.099	0.310	0.300	40
	FSL-RFO	**0.381**	**0.132**	**0.379**	0.300	40
	FSL-RFH	**0.381**	0.110	**0.379**	**0.350**	43

Finally, Table III reports the detailed recall performance of the best- and worst-performing aspects on MA-News, separately for each model and unlearning method on both the retain and forget splits. For each model, the best and worst aspects are determined based on the average performance across the zero-shot and ICU strategies considered. On the retain split, the aspect where all three models achieve the best performance is '*travel*', with Qwen obtaining the highest score. The most challenging retain aspect is '*science*' for Llama and '*news*' for Qwen and Phi. On the forget split, the best-performing aspect differs across models (i.e., '*science*', '*travel*', and '*entertainment*' for Llama, Qwen, and Phi, respectively), with each model achieving higher results under different ICU strategies. The most challenging forget aspect is '*sport*' for Llama and Qwen, and '*news*' for Phi.

We also report the number of occurrences of the '*health*' forget aspect in the generated summaries, which should be minimized. Except for Phi, which exhibits the best overall performance in the zero-shot setting (i.e., 28 occurrences), Qwen and Llama achieve their best results using the RFW and RFO strategies, respectively. This further highlights the beneficial effect of few-shot examples particularly on forget performance. Analyzing the MU baseline results, both GA (with only a few exceptions) and DPO achieve significantly higher scores compared to the zero-shot and ICU strategies across all best- and worst-performing aspects (e.g., Qwen 0.924 vs. 0.400 using the RFH strategy for '*travel*' on test retain). Except for Llama with DPO, the baselines also yield

the best performance on the forget aspect, resulting in the lowest numbers of occurrences (e.g., Phi 7 with GA and 6 with DPO).

C. Comparison of computational costs

Although MU baselines perform significantly better than the zero-shot and few-shot ICU strategies, they require a larger amount of resources. Before being deployed at inference time, they involve both an initial fine-tuning stage and the subsequent unlearning phase. Considering as an example the Phi model with the DPO approach, when analyzing only inference times, few-shot strategies with $k = 5$ are three times slower than the MU baselines. However, when fine-tuning and unlearning times are also taken into account, ICU strategies are nearly ten times faster than DPO. In terms of memory utilization, while few-shot strategies are slightly more expensive at inference time due to the inclusion of additional examples in the prompt, they do not require extra memory during training. Overall, ICU approaches are therefore about 1.5 times less memory-demanding than MU baselines. To summarize, although ICU strategies achieve worse retain and forget performance compared to more advanced and computationally expensive MU baselines, they represent a lightweight alternative in resource-constrained settings.

V. Conclusions and future works

This paper investigated the performance of In-Context Unlearning (ICU) approaches for text summarization, comparing them against classical Machine Unlearning (MU) methods on two benchmark datasets. ICU demonstrates greater flexibility and adaptability than MU, which requires ad hoc training and fine-tuning. ICU performance on text summarization is worse than that of state-of-the-art approaches (e.g., DPO) and comparable to baseline MU techniques (e.g., GA). In terms of complexity, ICU achieved a 10x speed-up in computational time and a 1.5x reduction in memory usage.

As future work, we plan to design and evaluate new ICU prompting strategies tailored to larger LLMs and new summarization tasks (e.g., temporal summarization [17], dialogue summarization [18]), as well as to explore their use for content moderation by forgetting undesired style attributes [19].

Acknowledgment

This study was partially carried out within the FAIR (Future Artificial Intelligence Research) and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1555.11-10-2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] B. Liu, Q. Liu, and P. Stone, "Continual learning and private unlearning," in *Proceedings of The 1st Conference on Lifelong Learning Agents*, ser. Proceedings of Machine Learning Research, vol. 199. PMLR, 22–24 Aug 2022, pp. 243–254. [Online]. Available: https://proceedings.mlr.press/v199/liu22a.html

[2] J. Geng, Q. Li, H. Woisetschlaeger, Z. Chen, Y. Wang, P. Nakov, H.-A. Jacobsen, and F. Karray, "A comprehensive survey of machine unlearning techniques for large language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.01854

[3] X. Feng, Y. Li, H. Ji, J. Zhang, L. Zhang, T. Du, and C. Chen, "Bridging the gap between preference alignment and machine unlearning," 2025. [Online]. Available: https://arxiv.org/abs/2504.06659

[4] M. Pawelczyk, S. Neel, and H. Lakkaraju, "In-context unlearning: language models as few-shot unlearners," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[5] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li, K. R. Varshney, M. Bansal, S. Koyejo, and Y. Liu, "Rethinking machine unlearning for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.08787

[6] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[7] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 048–11 064. [Online]. Available: https://aclanthology.org/2022.emnlp-main.759/

[8] B. Tan, L. Qin, E. P. Xing, and Z. Hu, "Summarizing text on any aspects: A knowledge-informed weakly-supervised approach," 2020. [Online]. Available: https://arxiv.org/abs/2010.06792

[9] H. Lin, J. Zhu, L. Xiang, F. Zhai, Y. Zhou, J. Zhang, and C. Zong, "Topic-oriented dialogue summarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1797–1810, 2023.

[10] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[11] P. Maini, Z. Feng, A. Schwarzschild, Z. C. Lipton, and J. Z. Kolter, "TOFU: A task of fictitious unlearning for LLMs," in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=B41hNBoWLo

[12] Llama Team, "Llama 3.2," 2024. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[13] Qwen Team, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[14] Microsoft, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," 2025. [Online]. Available: https://arxiv.org/abs/2503.01743

[15] J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo, "Knowledge unlearning for mitigating privacy risks in language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 389–14 408. [Online]. Available: https://aclanthology.org/2023.acl-long.805/

[16] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: your language model is secretly a reward model," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[17] D. R. Cambrin, L. Cagliero, and P. Garza, "DQNC2S: dqn-based cross-stream crisis event summarizer," in *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part III*, ser. Lecture Notes in Computer Science, vol. 14610. Springer, 2024, pp. 422–430. [Online]. Available: https://doi.org/10.1007/978-3-031-56063-7_34

[18] G. Gallipoli, L. Cagliero, and P. Garza, "Extractive conversation summarization driven by textual entailment prediction," in *2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT)*, 2023, pp. 1–6.

[19] M. La Quatra, G. Gallipoli, and L. Cagliero, "Self-supervised text style transfer using cycle-consistent adversarial networks," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 5, Nov. 2024. [Online]. Available: https://doi.org/10.1145/3678179