Exploring Cross-Client Memorization of Training Data in Large Language Models for Federated Learning

Tinnakit Udsa[1], Can Udomcharoenchaikit[1], Patomporn Payoungkhamdee[1], Sarana Nutanong[1], Norrathep Rattanavipanon[2]

[1]School of Information Science and Technology, VISTEC
[2]College of Computing, Prince of Songkla University
{tinnakit.u_s24, canu_pro, patomporn.p_s21, snutanon}@vistec.ac.th
norrathep.r@phuket.psu.ac.th

Abstract

Federated learning (FL) enables collaborative training without raw data sharing, but still risks training data memorization. Existing FL memorization detection techniques focus on one sample at a time, underestimating more subtle risks of cross-sample memorization. In contrast, recent work on centralized learning (CL) has introduced fine-grained methods to assess memorization across all samples in training data, but these assume centralized access to data and cannot be applied directly to FL. We bridge this gap by proposing a framework that quantifies both intra- and inter-client memorization in FL using fine-grained cross-sample memorization measurement across all clients. Based on this framework, we conduct two studies: (1) measuring subtle memorization across clients and (2) examining key factors that influence memorization, including decoding strategies, prefix length, and FL algorithms. Our findings reveal that FL models do memorize client data, particularly intra-client data, more than inter-client data, with memorization influenced by training and inferencing factors. Code for the framework is available at https://github.com/tinnakitudsa/FL_memorization_framework.git.

1 Introduction

Federated learning (FL) allows collaborative model training across multiple clients without sharing raw data, thereby preserving privacy in domains like healthcare. However, FL does not eliminate the risk of memorization, where large language models (LLMs) may inadvertently encode sensitive data.

Prior work employs techniques such as canary injection, verbatim and exact memorization, k-extractible metrics, and BLEU score (Carlini et al., 2019, 2022, 2021; Tirumala et al., 2022; Biderman et al., 2023; Ippolito et al., 2023; Kiyomaru et al., 2024) to measure memorization of LLMs in centralized learning (CL). However, these techniques share a critical assumption: a memorized text (suffix) can only be triggered by a prompt (prefix) from the same sample. This assumption does not hold in FL, where memorization may occur across clients (and thus across samples). Recent techniques (Zeng et al., 2024; Lee et al., 2023) lifted this assumption by proposing measurement of cross-sample memorization in CL.

In contrast, existing research on FL memorization of LLMs has focused primarily on canary injection – embedding out-of-distribution phrases into training data to see if the model reproduces them at inference time (Thakkar et al., 2021; Ramaswamy et al., 2020). While these techniques can detect verbatim in the same sample, they are poorly suited to capturing more realistic leakage across samples of actual in-distribution training data. Thus, FL memorization is likely to underestimate the real amount of memorization of training data.

This work is centered around the question: *How can we adapt cross-sample memorization assessment from CL to measure realistic memorization risks in FL, and what factors influence such leakage?* To address this question, we propose a framework that measures cross-client memorization in FL. The crux of our framework lies in a new *pairwise* technique that extends prior CL methods to estimate memorization between FL clients.

Based on this framework, we then formulate the following studies to answer the above question. First, we assess the extent to which the global model in an FL setup memorizes client training data in a cross-client fashion, capturing both intra- and inter-client memorization. Second, we empirically analyze the potential factors that may affect memorization, including decoding methods, prefix length, federated algorithm, model size, and the number of communication rounds. Together, these studies offer a comprehensive view of memorization risks across all clients in FL and show how CL insights can be adapted to decentralized contexts.

The main contributions of this study are outlined

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 679–693

as follows. **i) Problem Formulation:** We frame the challenge of detecting subtle memorization in FL by adapting cross-sample assessment methods from CL. **ii) Framework:** We design a framework that measures intra- and inter-client memorization in FL across clients, and conduct studies to quantify its extent and influencing factors. **iii) Key insights:** FL memorization occurs across clients, but its effect is less pronounced than within the same client. The memorization degree depends on the prefix length, the decoding strategy, and the federated algorithm. However, we did not observe a clear trend for model sizes and communication rounds. Additionally, the lengths of training inputs and outputs influence the length of generated text, which can limit the extractability of memorization in short-output tasks (e.g., classification). Finally, suffixes associated with certain clients appear to be more susceptible to memorization than those of others.

2 Memorization in Centralized Learning

We consider a CL setting in which a server trains an LLM M on a local dataset D consisting of N text samples, i.e., $D = \{d_i\}_{i=1}^{N}$, where d_i is split into a fixed-length prefix p_i and suffix s_i such that $d_i = p_i||s_i$. Let P and S denote the set of all prefixes and suffixes, i.e., $P = \{p_i\}_{i=1}^{N}$ and $S = \{s_i\}_{i=1}^{N}$, prior work (Zeng et al., 2024) defines memorization as:

Definition 2.1 In-distribution CL memorization A prefix $p \in P$ is said to induce memorization of a model M in its in-distribution training data if there exists suffix $s \in S$ such that $\mathcal{F}(M(p), s) =$ True for some discriminative function $\mathcal{F}$ that determines the similarity between two texts.

Then, MR, the memorization ratio of M on its training data, can be computed as the fraction of prefixes in P satisfying Definition 2.1.

In addition to cross-sample memorization, Zeng et al. (2024) adapts fine-grained memorization measurement that goes beyond exact matches. They instantiated $\mathcal{F}$ using the PAN2014 plagiarism detector (Sanchez-Perez et al., 2014, 2015) measuring text similarity at three levels of granularity: (i) **verbatim**, (ii) **paraphrase** and (iii) **idea-level**.

3 Proposed Study

3.1 Memorization in Federated Learning

In FL, we consider L clients, where each client C_i holds a private local dataset D_i. Similar to CL, each D_i can be divided into prefixes P_i and suffixes S_i. The aim in FL is to collaboratively train a shared global model M, without directly sharing any D_i. FL background is further provided in Appendix A.

To define in-distribution memorization in FL, we propose to generalize Definition 2.1 as follows:

Definition 3.1 In-distribution FL memorization A prefix $p_j \in P_j$ from client C_j is said to induce memorization of M on the in-distribution training data of client C_k if there exists suffix $s_k \in S_k$ such that $\mathcal{F}(M(p_j), s_k) =$ True for a discriminator $\mathcal{F}$.

We categorize FL memorization into two types. The first, *inter-client memorization*, occurs when $C_j \neq C_k$ (Figure 1B). This type is considered *harmful* since C_j can use its own prefix p_j to extract a suffix s_k belonging to another client C_k, leading to direct privacy leakage in FL. The second type, *intra-client memorization*, occurs locally on a client's own data, i.e., $C_j = C_k$ in Figure 1A. While this does not directly violate inter-client privacy, it may still pose a risk if the client's prefixes are known to others (e.g., P_j is public). As such, we classify it as *harm-exposed* rather than harmful.

Memorization Metrics. FL memorization, unlike CL, occurs across clients. We define a *pairwise memorization ratio* ($\mathsf{MR}_{j\to k}$) from client C_j to C_k as a fraction of prefixes from C_j causing the model M to memorize a suffix belonging to C_k:

$$\mathsf{MR}_{j\to k} = \frac{|P_{j,k}|}{|P_j|} \tag{1}$$

where $P_{j,k}$ denotes a subset in P_j that induce memorization on C_k w.r.t. Definition 3.1.

Using $\mathsf{MR}_{j\to k}$, we define $\mathsf{MR}_{\mathsf{Intra}}$ and $\mathsf{MR}_{\mathsf{Inter}}$ as the memorization ratios for intra-client and inter-client memorizations, respectively, averaged across all client pairs. Intuitively, they reflect how often a prefix from one client causes the model to memorize a private suffix of either the same ($\mathsf{MR}_{\mathsf{Intra}}$) or a different ($\mathsf{MR}_{\mathsf{Inter}}$) client. To compare CL vs FL memorization, we introduce $\mathsf{MR}_{\mathsf{TotalCL}}$ and $\mathsf{MR}_{\mathsf{TotalFL}}$, which capture the *total* number (in percentage) of memorization-inducing prefixes from all clients in CL and FL, respectively. Appendix D provides formal definitions of these metrics.

Estimating FL Memorization. Precise measurement of memorization in LLMs is a challenging task (Carlini et al., 2022). To tackle this, we extend the estimation method proposed by Zeng et al. (2024); Lee et al. (2023) to compute $\mathsf{MR}_{j\to k}$. Given prefixes P_j from client C_j and suffixes S_k from C_k, our framework calculates $\mathsf{MR}_{j\to k}$ by estimating P_j fraction that induces memorization on

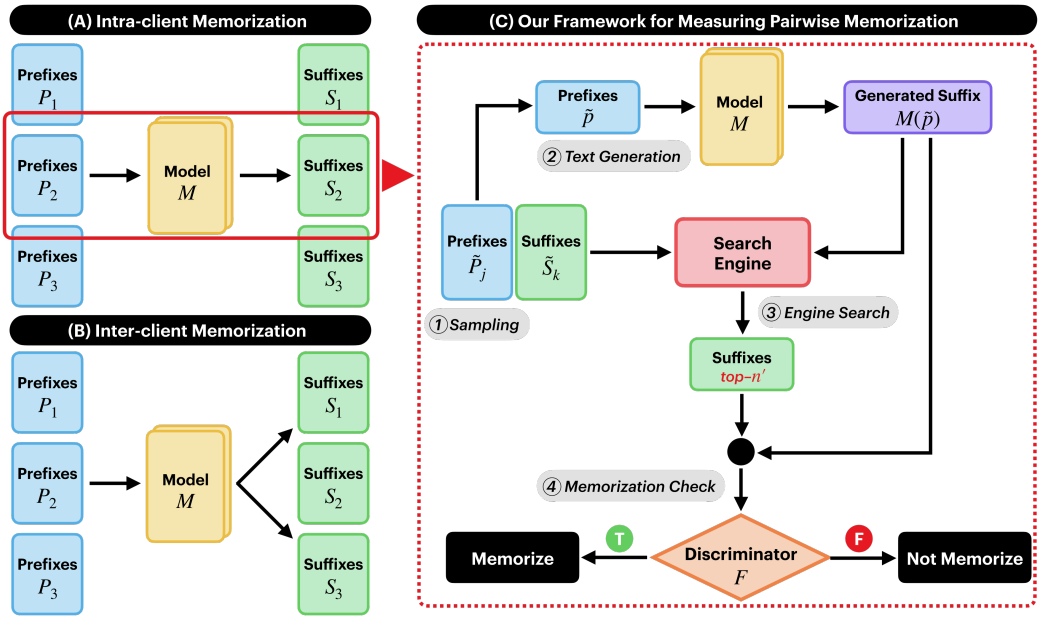

Figure 1: Harm-exposed memorization (A) uses same-client prefixes/suffixes as input to our framework (C) (adapted from Lee et al. (2023); Zeng et al. (2024)) while harmful memorization (B) uses those from different clients.

S_k. Figure 1 shows the steps in our framework.

In ①, we sample $\tilde{P}_j \subset P_j$ and $\tilde{S}_k \subset S_k$ of size $n \ll \min(|P_j|, |S_k|)$, where $n = 4K$ in our experiments. We use each prefix $\tilde{p} \in \tilde{P}_j$ as a prompt to M, generating a text $M(\tilde{p})$ in ②. In ③, we index $\tilde{S}_k$ using Elasticsearch, and query it with each $M(\tilde{p})$ to retrieve top-n' suffixes from $\tilde{S}_k$ that are most similar to $M(\tilde{p})$; following Zeng et al. (2024), $n' = 10$. In ④, we compare $M(\tilde{p})$ with the top-n' suffixes using the PAN2014 plagiarism detector, which acts as the discriminative function $\mathcal{F}$. Finally, in ⑤, if the detector returns *True* (from any of the verbatim, paraphrase or idea-level similarity types) for any top-n' suffixes, we consider $\tilde{p}$ to induce memorization w.r.t. Definition 3.1 and estimate $\mathsf{MR}_{j \to k}$ as the fraction of prefixes in $\tilde{P}_j$ triggering *True*. Based on $\mathsf{MR}_{j \to k}$, aforementioned memorization metrics can be derived (see Appendix D). As a result, these metrics let us explore the following RQs.

3.2 Research Questions (RQs)

RQ1: Do FL models memorize training data?

In this RQ, we aim to empirically evaluate $\mathsf{MR}_{\mathsf{Intra}}$ and $\mathsf{MR}_{\mathsf{Inter}}$ for LLMs trained under FL settings to better understand the associated privacy risks. As byproducts, we also examine: (1) whether FL leads to higher/lower memorization compared to CL, i.e., whether $\mathsf{MR}_{\mathsf{TotalCL}} > \mathsf{MR}_{\mathsf{TotalFL}}$ and (2) how model accuracy influences memorization.

RQ2: Which factors impact FL memorization?

As shown in prior studies (Zeng et al., 2024; Lee et al., 2023; Kiyomaru et al., 2024; Carlini et al., 2022), the extent of CL memorization can be attributed to several factors. Shifting from CL to FL, we aim to investigate whether similar patterns hold. To this end, we perform ablation studies to examine the impact of both existing factors and FL-specific factors on FL memorization.

3.3 Experimental Setup

Datasets. We evaluate our approach on four tasks: summarization, dialog, question answering (QA), and classification. For each task, we consider 3 FL clients, where each client contains a distinct dataset originating from a domain with potential risks of sensitive information leakage. Detailed descriptions are provided in Appendix B.

Models. Our main experiments are conducted using Qwen2.5-3B (Qwen et al., 2025). In addition, we show the results for Llama3.2 (Grattafiori et al., 2024) and GPT2 (Radford et al., 2019), along with further results of model size variations, in Appendix G.

Default Setting. Unless stated otherwise, we report the results using the following setup: Qwen2.5-3B with prefix length of 30, top-k decoding method, and 3 FL rounds trained under FedAvg (McMahan et al., 2017). Detailed configurations are shown in Appendix C.

4 Results

4.1 Memorization in Federated Learning

As shown in Table 1, the models do memorize training data even in FL settings. In particular, MR_{Intra} is consistently higher than MR_{Inter} across all evaluated tasks. This answers *RQ1* that FL memorization tends to occur more within the same client than across clients. Interestingly, the classification task shows no memorization. Further analysis of this behavior, including correlations between input and output lengths, is provided in Section 5.

Table 1: Intra- vs inter-client memorization in FL

Tasks	MR_{Intra} (%)	MR_{Inter} (%)
Summarization	0.342	0.046
Dialog	1.533	1.446
QA	1.450	0.813
Classification	0.000	0.000

Table 2 shows no clear trend in memorization when shifting from CL to FL. This finding contrasts with Thakkar et al. (2021), which reported reduced memorization in FL. We expect this difference arises from the setup of memorization measurements: their work evaluates out-of-distribution memorization (via canary injection), whereas ours focuses on the in-distribution one. Additionally, we found no correlation between memorization and performance in either setting.

Table 2: Trade-off between CL and FedAvg

Task	Models	Performance	Memorization(%)
Summarization	$MR_{TotalCL}$	28.46	0.558
	$MR_{TotalFL}$	29.88	0.433
Dialog	$MR_{TotalCL}$	19.40	3.417
	$MR_{TotalFL}$	18.11	3.992
QA	$MR_{TotalCL}$	26.66	2.150
	$MR_{TotalFL}$	28.60	2.917
Classification	$MR_{TotalCL}$	76.30	0.000
	$MR_{TotalFL}$	51.22	0.000

Llama3.2-3B shows similar results to Qwen2.5-3B; see Appendix G.1 and G.2 for detailed results.

4.2 Memorization Factors

Next, to answer *RQ2*, we investigate the potential factors that influence MR_{Intra} and MR_{Inter}.

Decoding Method. Table 3 indicates memorization tends to increase when applying top-k or top-p decoding. The results align with prior work (Zeng et al., 2024; Lee et al., 2023), where sophisticated decoding strategies amplify memorization effects.

Prefix Length. As shown in Table 4, increasing prefix length lowers MR_{Inter} and MR_{Intra} in most cases, indicating that shorter prefixes can induce greater memorization in FL.

Table 3: Memorization with various decoding methods

Tasks	Decoding	MR_{Intra} (%)	MR_{Inter} (%)
Summarization	temperature	0.475	0.067
	top-k	0.342	0.046
	top-p	0.525	0.050
Dialog	temperature	1.267	1.442
	top-k	1.533	1.446
	top-p	3.792	2.996
QA	temperature	1.283	0.750
	top-k	1.450	0.813
	top-p	2.567	1.438

Table 4: Memorization with various prefix lengths

Tasks	Prefix Length	MR_{Intra} (%)	MR_{Inter} (%)
Summarization	10	0.508	0.188
	30	0.342	0.046
	50	0.425	0.038
	100	0.208	0.004
Dialog	10	2.108	1.992
	30	1.533	1.446
	50	1.575	1.242
	100	1.408	1.150
QA	10	1.525	1.383
	30	1.450	0.813
	50	1.242	0.550
	100	1.125	0.429

Federated Algorithm. Table 5 indicates that FedProx leads to higher memorization rates (MR_{Inter} and MR_{Intra}). Notably, it exhibits memorization in the classification task, which is absent in FedAvg. This indicates that memorization varies across algorithms and should be evaluated independently.

Table 5: Memorization across federated algorithms

Tasks	Federated Algorithm	MR_{Intra} (%)	MR_{Inter} (%)
Summarization	FedAvg	0.342	0.046
	FedProx	0.942	0.138
Dialog	FedAvg	1.533	1.446
	FedProx	1.892	1.879
QA	FedAvg	1.450	0.813
	FedProx	3.675	2.146
Classification	FedAvg	0.000	0.000
	FedProx	0.011	0.000

Besides the aforementioned factors, our experiments also consider model size and the communication rounds. The results show no strong association with MR_{Inter} and MR_{Intra}; detailed results are provided in Appendix G.6 and G.7.

5 Observations on Memorization Patterns

Here, we conduct two analyses to gain deeper understanding of memorization in FL.

5.1 Correlations Between Input Length, Output Length, and Generated Suffix Length

Table 6: Token length of input, output, and generated text. (Median)

Model	Task	Input	Output	Generated Text
Qwen2.5-3B	Summarization	184	15	10
	Dialog	86	108	59
	QA	325	47	78
	Classification	38	2	1
Llama3.2-3B	Summarization	182	15	6
	Dialog	84	107	66
	QA	313	47	26
	Classification	37	2	2

First, we analyze the correlations between training input length, training output length, and generated suffix length (from our memorization framework). Table 6 shows that the generated suffix length is strongly affected by the input and output lengths Specifically, in summarization and classification tasks, the median of generated suffix lengths is close to the output length. This phenomenon leads the classification task to generate a smaller token than other tasks to the point that its generated suffix length is smaller than the memorization threshold set in PAN2014, leading to 0.000% memorization in most of our experiments.

5.2 Effects of Suffixes

Table 7: Effect of Suffix Index(%)

Prefix	Suffix	$\mathsf{MR}_{j\to k}$
Group1	Group1	1.450
Group1	Group2	1.525
Group1	Group3	1.500
Group2	Group1	1.150
Group2	Group2	1.200
Group2	Group3	1.225
Group3	Group1	1.725
Group3	Group2	1.550
Group3	Group3	1.950

It is worth noting from Table 7 that memorization is more likely when the prefix and suffix are from the same client. Surprisingly, suffixes from Group3 of Dialog task (see Appendix B.2) are memorized more than those from other clients. This suggests that the characteristic of a dataset to which suffixes belong plays an important role in their chance of being memorized. Future research is needed to better understand which specific dataset properties drive this effect and how they interact with FL setting.

6 Conclusion

We presented a framework for evaluating fine-grained cross-sample and cross-client memorization in FL, adapting techniques from CL to measure both intra- and inter-client leakage. Empirical results show that FL models tend to memorize training data, with the intra-client memorization consistently higher than the inter-client memorization in all cases. Among all factors examined, our findings indicate that FL memorization rates depend on prefix length, decoding strategy, and federated algorithm. Also, we find no clear evidence that memorization is reduced in FL compared to CL, emphasizing the need to quantify privacy leakage even when privacy-sensitive applications are trained under FL.

Limitations

Despite its contributions, this study has limitations. Firstly, the PAN2014 plagiarism detector can produce misleading results when models generate incoherent output (see Appendix E.5). Note that, for the results reported in our paper, we use models that can generate coherent outputs. Secondly, the PAN2014 plagiarism detector, can only measures fine-grained memorization in English texts. Lastly, while this study provides empirical results, theoretical investigation into why intra-client memorization exceeds inter-client memorization is an interesting research direction.

Acknowledgments

Computing resources were supported part by the ThaiLLM collaborations, funded by the Digital Economy and Society Development Fund of the Ministry of Digital Economy and Society, Thailand; the Fundamental Fund (FF 87245) from Thailand Science Research and Innovation (TSRI), 2025; and the WangchanX project through donations from SCB, SCBX, and PTT. Appreciation is extended to members of the Natural Language Processing and Representation Learning Lab at VISTEC for their technical and moral support, especially Wuttikorn Ponwitayarat, Pume Tuchinda, and Natnasa Lertmahakul. The first author thanks the co-authors and co-advisors for their guidance and insightful feedback throughout this work, friends and family for their encouragement and understanding, and himself for the perseverance and dedication that made this work possible.

References

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA. USENIX Association.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Geiger, R. Stuart. 2020. Arxiv archive: A tidy and complete archive of metadata for papers on arxiv.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan. Association for Computational Linguistics.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3637–3647, New York, NY, USA. Association for Computing Machinery.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Preprint*, arXiv:1812.06127.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Jaehoon Oh, SangMook Kim, and Se-Young Yun. 2022. FedBABU: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Swaroop Ramaswamy, Om Thakkar, Rajiv Mathews, Galen Andrew, H. Brendan McMahan, and Françoise Beaufays. 2020. Training production language models without memorizing user data. *Preprint*, arXiv:2009.10031.

M. A. Sanchez-Perez, A. Gelbukh, and G. Sidorov. 2015. Adaptive algorithm for plagiarism detection: The best-performing approach at pan 2014 text alignment competition. In *Lecture Notes in Computer Science*, volume 9283, pages 402–413. Springer.

Miguel Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh. 2014. The winning approach to text alignment for text reuse detection at pan 2014. In

CLEF2014 Working Notes, volume 1180 of *CEUR Workshop Proceedings*, pages 1004–1011, Sheffield, UK. CEUR-WS.org. Notebook for PAN at CLEF 2014.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Terminology-aware medical dialogue generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Om Dipakbhai Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Francoise Beaufays. 2021. Understanding unintended memorization in language models under federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 1–10, Online. Association for Computational Linguistics.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: analyzing the training dynamics of large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2024. Exploring memorization in fine-tuned language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3917–3948, Bangkok, Thailand. Association for Computational Linguistics.

Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Fedala: adaptive local aggregation for personalized federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Appendix

A Federated Learning

The goal of FL is to collaboratively train a model from multiple sources without directly sharing their data. The widely adopted FL framework is FedAvg (McMahan et al., 2017). In FedAvg, each client trains a local model on its own dataset. Afterward, it sends the locally trained model to a central server where all local models are aggregated via averaging. The averaged model is sent back to local clients to continue training, repeating this process iteratively. Ultimately, the globally averaged model is used for all clients.

Another FL algorithm used in this work is FedProx (Li et al., 2020). It addresses the heterogeneity issue of FedAVG (e.g., non-IID data) by modifying the local objective of each client to include a proximal (regularization) term that penalizes deviations from the current global model.

B Datasets

To investigate the impact of fine-tuning in FL, we conduct separate experiments with publicly available datasets on summarization, dialog, question-answering, and classification tasks. Each dataset is partitioned among clients in a non-IID fashion. Task-specific details are provided below.

B.1 Summarization

Following Lee et al. (2023), we use the ArxivAbstract (Geiger, R. Stuart, 2020) dataset, as too much memorization in academic writing can be seen as a form of plagiarism. The dataset contains abstracts collected from arXiv.org. We focus on three subject areas: Astrophysics, Condensed Matter, and Mathematics. For each area, we sample 30,000 examples, with 27,000 used for training and 3,000 for testing. The subcategories are as follows:

i) **Astrophysics:** High Energy Astrophysical Phenomena, Instrumentation and Methods for Astrophysics, Earth and Planetary Astrophysics, and Astrophysics of Galaxies.
ii) **Condensed Matter:** Superconductivity, Soft Condensed Matter, Disordered Systems and Neural Networks, Quantum Gases, and Other Condensed Matter.
iii) **Mathematics:** K-Theory and Homology, Statistics Theory, Differential Geometry, Mathematical Physics.

This dataset is publicly available at https://github.com/staeiou/arxiv_archive.

B.2 Dialog

We use the data from Tang et al. (2023), consisting of 100,000 real patient-doctor dialogues sourced from HealthCareMagic.com. We divide the dataset into 3 disjoint groups. For each group, we sample 30,000 instances with 27K/3K for the train-test split and assign them to each FL client. The groups are defined as follows:

i) **Group 1:** Child Health, Lung and Chest disorders, Mental Health, Cancer, Birth Control, Bones, Muscles and Joints, Natural and Home Remedies, and Accident and Emergency.
ii) **Group 2:** Brain and Spine, Infections, Abdominal Pain, Liver and Gall Bladder, Kidney Conditions, Infertility Problem, Asthma and Allergy, Diabetes, and Lupus.
iii) **Group 3:** Men's Health, Hypertension and Heart Disease, Dental Health, Lump, Pregnancy, Pain Management, Medicines and Side Effects, and Alternative Medicine.

The dataset can be found in https://github.com/tangg555/meddialog.

B.3 Question-Answering

We choose PubMedQA (Jin et al., 2019), the medical domain dataset, as the risk of sensitive information leakage is particularly critical in this field. The dataset contains over 273,000 biomedical question-answering instances, including 1,000 expert-labeled, 61,200 unlabeled, and 211,300 artificially generated samples, all derived from PubMed abstracts. The task involves answering questions with one of three labels: yes, no, or maybe. Each instance includes both a short answer (the label) and a long-form explanation. In our experiments, we use the long answers as the ground truth labels. The dataset is available at: https://huggingface.co/datasets/bigbio/pubmed_qa.

We divide the dataset into 3 groups based on Medical Subject Headings. Each clients contain 30,000 sampled instances, with 27,000 used for training and 3,000 for testing. The group details are as follows:

i) **Group 1:** having no Humans tag and having Animals tag.
ii) **Group 2:** having no Animals tag, having Humans tag, and having Middle Aged tag.
iii) **Group 3:** having no Animals tag, having Humans tag, and having no Middle Aged tag.

Table 8: The chat templates for each task employed in fine-tuning the LLMs and in subsequent evaluation.

Summarization
User: Please summarize the following abstract into a title. {Abstract} **Assistant:** {Title}
Dialog
User: If you are a doctor, please answer the medical questions based on the patient's description. {Patient} **Assistant:** {Doctor}
Question Answering (QA)
User: {Question} {Context} **Assistant:** {Answer}
Classification
User: Please classify the following passage into one of the following categories: BACKGROUND, OBJECTIVE, METHODS, RESULTS, or CONCLUSIONS. {Passage} **Assistant:** {Class}

B.4 Classification

We use the PubMed 200k RCT (Dernoncourt and Lee, 2017) dataset, derived from PubMed, for sequential sentence classification. It contains around 200,000 abstracts from randomized controlled trials. It consists of 5 components: background, objective, methods, results, and conclusion. This dataset is publicly available at `https://github.com/Franck-Dernoncourt/pubmed-rct`. Following standard practice of data partition for classification task in FL (Oh et al., 2022; Zhang et al., 2023), we partition the dataset among FL clients using a Dirichlet distribution with $\alpha = 5.0$, where each client receives 30K instances with 27K/3K train-test split.

B.5 Sample Format

The pre-processed data used for model training are formatted according to Table 8. The formatted data are used to train the model with its corresponding template applied.

C Training Costs and Hyperparameters

C.1 Training Cost

We used an NVIDIA A100 GPU with 4,000 GPU hours to train Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Llama-3.2-1B, and Llama-3.2-3B. It takes the total of 1,600 GPU hours and 1,600 CPU hours to apply our memorization framework across all experiments.

C.2 Hyperparameters

We trained models using LLaMA Factory library[1] with a 2e-4 learning rate, bfloat16 and batch sizes of 64.

C.3 Memorization Factors

Decode Method: Following Lee et al. (2023), we empirically study multiple decoding methods: top-k, top-p, and temperature. For each decoding, we start with the default Huggingface trainer-generated parameters and update them with the following: `k=40` for top-k, `p=0.8` for top-p, and `temperature=1.0` for temperature decoding.

Prefix Length: Our experiments are conducted with different prefix lengths: 10, 30, 50, and 100 tokens, following Zeng et al. (2024).

Federated Algorithm: We use FedAvg and FedProx as federated algorithm to train and measure memorization.

Communication Rounds: We conduct experiments using 1, 3, and 5 federated communication rounds, which effectively correspond to the number of training epochs in the FL process.

[1] `https://github.com/hiyouga/LLaMA-Factory`

Model Size: We select Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Llama-3.2-1B, adn Llama-3.2-3B, which have 494M, 1.54B, 3.09B, 1.24B, and 3.21B parameters, respectively, to measure memorization.

D Formal Definitions

D.1 Harm-exposed and Harmful Memorization Ratios

Based on the notion of *intra-client (harm-exposed)* memorization and Equation 1, we can define $\mathsf{MR}_{\mathsf{Intra}}$, the overall harm-exposed memorization ratio, as a weighted sum on all L clients:

$$\mathsf{MR}_{\mathsf{Intra}} = \sum_{j=1}^{L} w_j \cdot \mathsf{MR}_{j \to j} \tag{2}$$

where w_j represents the weight of client C_j; in our work, it is calculated as the ratio of C_j's training dataset to the total training data among all L clients, i.e., $w_j = |D_j| / \sum_{i=1}^{L} |D_i|$.

Then, we can compute the average harmful memorization ratio incurred by a specific client C_j as:

$$\mathsf{MR}_{\mathsf{Inter}}(j) = \frac{1}{L-1} \sum_{j \neq k} \mathsf{MR}_{j \to k} \tag{3}$$

and finally $\mathsf{MR}_{\mathsf{Inter}}$, the overall harmful memorization ratio, across all clients as a weighted sum:

$$\mathsf{MR}_{\mathsf{Inter}} = \sum_{j=1}^{L} w_j \cdot \mathsf{MR}_{\mathsf{Inter}}(j) \tag{4}$$

In our experiments, we approximate $\mathsf{MR}_{\mathsf{Intra}}$ and $\mathsf{MR}_{\mathsf{Inter}}$ following the estimated $\mathsf{MR}_{j \to k}$ produced from the framework in Section 3.1.

D.2 Total Memorization Ratios in CL and FL

To enable a fair comparison between CL and FL, we use a metric based on the *total* number of prefixes that cause the model to memorize *any* suffix. In FL, this number corresponds to the union[2] of all $P_{j,k}$-s. Consequently, we define the total memorization ratio in FL as:

$$\mathsf{MR}_{\mathsf{TotalFL}} = \frac{|\bigcup_{j,k} P_{j,k}|}{|\bigcup_{j} P_j|} \tag{5}$$

The total memorization ratio in CL, $\mathsf{MR}_{\mathsf{TotalCL}}$, is computed the same way as $\mathsf{MR}_{\mathsf{TotalFL}}$ via Equation 5. The key distinction, however, lies in how the model M is trained: in CL, M is trained centrally on the combined datasets by a trusted third party, whereas in the FL setting, M is trained in a distributed manner across clients.

[2]Note that we take a *union* rather than a summation to avoid double-counting memorization-inducing prefixes as a single prefix may cause the model to leak suffixes belonging to multiple clients, i.e., it is possible that $P_{j,k1} \cap P_{j,k2} \neq \varnothing$ for $k1 \neq k2$.

E PAN2014 and Three Categories of Memorization

E.1 PAN2014 Plagiarism Detector

PAN2014 plagiarism detector evaluates text similarity across three distinct categories: verbatim, paraphrase, and idea. This allows for fine-grained assessment of memorization that extends beyond exact matches captured by verbatim similarity. For this work, we use the improved version of PAN2014 proposed in (Lee et al., 2023). Detailed procedures for measuring cross-sample memorization using PAN2014 can be found in Lee et al. (2023); Zeng et al. (2024).

E.2 Memorization Category

In this work, we use all 3 PAN2014 categories in our analysis. In PAN2014, paraphrasing is evaluated using RoBERTa and NER models, with predictions categorized as low-confidence when p < 0.5 and high-confidence when p > 0.5; as in Zeng et al. (2024), we report results for both categories in the breakdown provided in Appendix G.

E.3 Memorization Measurement

To measure memorization, we use the abstract, patient, context, and passage texts (see Table 8) as input (prefix+suffix) to our framework for the summarization, dialog, question-answering, and classification tasks, respectively.

E.4 Hyperparameters for Memorization Measurement

Following Zeng et al. (2024), we set the minimal match threshold to at least 50 characters for PAN2014 in all memorization types. We also filter out some odd cases, as described in Appendix E.5, using a naive filter that excludes samples with at least ten repetitions of a three-word sequence.

E.5 Memorization Result of Incoherent Output Generation

When a model generates incoherent outputs (see Table 9), the PAN2014 plagiarism detector misclassifies them as idea memorization. The reason for this phenomenon is that PAN2014 is designed to discriminate human-like texts, not incoherent texts.

Table 9: Misclassify Memorization of Incoherent Output Generation

Type	Machine-Written Text	Training Text
Idea	Algebraic Groups for Lie groups, Lie Groups and subalgebras, Lie algebras c to Lie algebras and Lie algebras such algebras, cohomologies, algebraic groups, Lie groups and Lie algebras without equations	Here we focus on contractions of Lie algebras and algebraic groups.
Idea	Your thyroid is a butterfly-shaped gland at the base of your throat. You have two lobes, lobes, and three lobes. It has two lobes, two lobes, one lobe, and two lobes, one lobe, and two lobes, two lobes, and two lobes, both lobes at the base of your throat. It's involved in some very big jobs, like managing your heart rate, blood pressure, body temperature, and weight. Your thyroid has two lobes, lobes, lobes, lobes, lobes, lobes, lobes, lobes, lobes, lobes, and weight.	the thyroid gland is removed. A lobectomy is when one of the two lobes of your thyroid is removed.

As this assumption does not hold, it can leads to unpredictable results. This seems to be an apparent limitation for any PAN2014-based memorization methods.

F Memorized Examples

Examples of memorization in the categories of verbatim, paraphrase, and idea are shown in Table 10.

G Detailed Results

G.1 Per-Category Results for $\mathsf{MR}_{\mathsf{Intra}}$ and $\mathsf{MR}_{\mathsf{Inter}}$

We report in Table 11 $\mathsf{MR}_{\mathsf{Intra}}$ and $\mathsf{MR}_{\mathsf{Inter}}$ results for each individual category of PAN2014 when used as the discriminator $\mathcal{F}$. This provides the breakdown results from our main table (Table 1). The results show that $\mathsf{MR}_{\mathsf{Intra}}$ is higher than $\mathsf{MR}_{\mathsf{Inter}}$, for most categories for Qwen2.5-3B, Llama-3.2-3B, and GPT-2 XL.

G.2 Trade-off Results between $\mathsf{MR}_{\mathsf{TotalCL}}$ and $\mathsf{MR}_{\mathsf{TotalFL}}$

We provide the breakdown results from Table 2 in Table 12.

G.3 Memorization with Various Decoding Methods

$\mathsf{MR}_{\mathsf{Intra}}$ and $\mathsf{MR}_{\mathsf{Inter}}$ tend to increase when top-p or top-k is used in most cases for both Qwen2.5-3B and Llama3.2-3B as shown in Table 13, suggesting better decoding methods lead to more memorization.

G.4 Memorization with Various Prefix Lengths

Table 14 shows that increasing prefix length lowers $\mathsf{MR}_{\mathsf{Inter}}$ and $\mathsf{MR}_{\mathsf{Intra}}$ in most cases for both Qwen2.5-3B and Llama3.2-3B. This phenomenon suggests that shorter prefixes can induce more memorization in FL.

G.5 Memorization across Federated Algorithms

FedProx has more $\mathsf{MR}_{\mathsf{Inter}}$ and $\mathsf{MR}_{\mathsf{Intra}}$ than FedAvg as shown in Table 15. It also results in memorization in the classification task which is not observed before in FedAvg for both Qwen2.5-3B and Llama3.2-3B.

G.6 Memorization with various Model Size

As shown in Table 16, we observe no apparent impact of model sizes of both Qwen2.5-3B and Llama3.2-3B on any memorization types across tasks.

G.7 Memorization with various Communication Rounds

Table 17 indicates that communication rounds does not noticeably affect memorization for either Qwen2.5-3B or Llama3.2-3B across tasks.

G.8 Trade-off between Performance and Memorization CL and FedAvg

We observe no clear trend in memorization when shifting from CL to FL for Qwen2.5-3B or Llama3.2-3B as shown in Table 18. Furthermore, we found no correlation between memorization and performance in either setting.

Table 10: Examples of memorization from FL

Type	Machine-Written Text	Training Text
Verbatim	... I am concerned about my high blood pressure, high cholesterol, and triglycerides. If they are high, what should be done?I can be managed by lifestyle changes, medications, and diet changes ...	... I have been diagnosed with interstitial cystitis, fibromyalgia, mild psoriasis, high blood pressure, high cholesterol and triglycerides and have frequent UTI's. ... (Dialog, prefix: group1, suffix: group2)
Verbatim	in development of the mitogen-activated protein kinase (MAPK) pathway. Reducing the expression of the Redd-1 protein may increase the levels of MAPK. In addition, increasing the expression of the Redd-1 protein may inhibit the progression of the MAPK pathway. The Redd-1 protein is known ...	... We characterized the signaling properties of confirmed molecular alterations by ectopic expression of engineered cDNAs in 293H cells. Activation of the mitogen-activated protein kinase (MAPK) pathway in cells by ectopic expression of PAPSS1-BRAF was abrogated by mitogen-activated protein ... (QA, prefix: group1, suffix: group3)
Verbatim	comparison with a ground-based system for the medium-sized telescopes of the Cherenkov Telescope Array	construction is scheduled to begin in fall at the Fred Lawrence Whipple Observatory in southern Arizona, USA. The Schwarzschild-Couder telescope is a candidate for the medium-sized telescopes of the Cherenkov Telescope Array, which utilizes ... (Summarization, prefix: astro, suffix: astro)
Paraphrase	do. I have a strong medical history of cancer to my family. There are chances, but we shouldn't worry much as these are likely benign tumors	of cancer, colon in 1996 and lung in 2005. I also have a strong history of cancer in my immediate family. While at the dentist, I noticed a skin tag ... (Dialog, prefix: group3, suffix: group1)
Paraphrase	PE) cellular functions including migration and apoptosis. We have developed a novel approach to study matrix effects on RPE, which provides the rationale for the treatment of many macular degenerative diseases through blocking active MMP. In this study, we investigated MMP effects on different cellular functions of RPE.	Cinaciguat, the novel soluble guanylate cyclase activator, currently being in phase IIb clinical trial, has been shown to exert antiplatelet and anti-remodeling effects in animal models of vascular pathology. In this study we investigated the effects of cinaciguat on post-injury arterial stenosis. Male Sprague-Dawley rats (n=100) underwent endothelial denudation ... (QA, prefix: group3, suffix: group1)
Paraphrase	(Target of Opportunity) missions to study the blazar source in depth. Fermi Large Area Telescope Target of Opportunity Missions	The data were taken with the Large Area Telescope on board the Fermi Gamma-ray Space Telescope. An extended source is found at a position consistent with that of RCW 103, and its emission was only detected above 1 GeV ... (Summarization, prefix: astro, suffix: astro)
Idea	am having very heavy throat pain and sometimes pain is felt in the chest , and sometimes it feels like something is pressing in the chest .i have cough and chest pain after eating rice . i can feel the lump in the throat near tonsil and it is also painful ...	clear from my throat. I also have a very sore throat and pain in my chest. Is this a viral infection I just have to wait out or could there be something I can do about it? (Dialog, prefix: group1, suffix: group2)
Idea	HEK293 are the four cell lines analyzed. The expression of mRNA or protein can be measured by quantitative PCR. The expression level of rRNA (total rRNA) and total mRNA (mRNA) may represent the amount of TmRNA present in the cells. The rRNA expression level of mRNA in L929 cells could be measured by quantitative PCR, but not by quantitative RT (real time PCR) method. ...	knocked down by small interfering RNA (siRNA) in HeLa cells which were then cultured in conventional medium or serum starvation medium. The protein level of TXNDC5 was evaluated by Western blot analysis. The mRNA level of TXNDC5 was measured by quantitative real-time PCR. Cell growth rate was determined by cell proliferation assay kit (MTS method). Cell cycle distribution and apoptosis were detected by flow cytometry. ... (QA, prefix: group1, suffix: group3)

Table 11: Per-Category Memorization in FL(%)

Model	Tasks	Type	Total	Verbatim	Idea	Paraphrase (p > 0.5)	Paraphrase (p < 0.5)
Qwen2.5-3B	Summarization	MR_{Intra}	0.342	0.008	0.000	0.025	0.308
		MR_{Inter}	0.046	0.000	0.000	0.013	0.033
	Dialog	MR_{Intra}	1.533	0.000	0.067	0.258	1.217
		MR_{Inter}	1.446	0.000	0.042	0.292	1.125
	QA	MR_{Intra}	1.450	0.000	0.067	0.375	1.033
		MR_{Inter}	0.813	0.000	0.046	0.192	0.588
	Classification	MR_{Intra}	0.000	0.000	0.000	0.000	0.000
		MR_{Inter}	0.000	0.000	0.000	0.000	0.000
Llama-3.2-3B	Summarization	MR_{Intra}	0.700	0.008	0.008	0.125	0.558
		MR_{Inter}	0.067	0.000	0.000	0.013	0.054
	Dialog	MR_{Intra}	3.167	0.000	0.183	0.692	2.367
		MR_{Inter}	2.313	0.000	0.071	0.488	1.813
	QA	MR_{Intra}	2.183	0.083	0.050	0.583	1.500
		MR_{Inter}	1.375	0.008	0.046	0.367	0.983
	Classification	MR_{Intra}	0.000	0.000	0.000	0.000	0.000
		MR_{Inter}	0.000	0.000	0.000	0.000	0.000
GPT-2 XL	Summarization	MR_{Intra}	0.567	0.008	0.067	0.075	0.417
		MR_{Inter}	0.075	0.000	0.017	0.008	0.050
	Dialog	MR_{Inter}	1.333	0.008	0.033	0.233	1.083
		MR_{Intra}	1.146	0.000	0.033	0.179	0.946
	QA	MR_{Intra}	0.958	0.008	0.033	0.192	0.733
		MR_{Inter}	0.558	0.000	0.008	0.125	0.438
	Classification	MR_{Intra}	0.000	0.000	0.000	0.000	0.000
		MR_{Inter}	0.000	0.000	0.000	0.000	0.000

Table 12: Per-Category Memorization Comparison between CL and FL(%)

Model	Tasks	Type	Verbatim	Idea	Paraphrase (p > 0.5)	Paraphrase (p < 0.5)
Qwen2.5-3B	Summarization	$MR_{TotalCL}$	0.008	0.033	0.042	0.475
		$MR_{TotalFL}$	0.008	0.000	0.050	0.375
	Dialog	$MR_{TotalCL}$	0.000	0.133	0.650	2.733
		$MR_{TotalFL}$	0.000	0.142	0.817	3.217
	QA	$MR_{TotalCL}$	0.050	0.208	0.517	1.483
		$MR_{TotalFL}$	0.000	0.158	0.742	2.092
	Classification	$MR_{TotalCL}$	0.000	0.000	0.000	0.000
		$MR_{TotalFL}$	0.000	0.000	0.000	0.000
Llama3.2-3B	Summarization	$MR_{TotalCL}$	0.017	0.025	0.108	0.792
		$MR_{TotalFL}$	0.008	0.008	0.150	0.642
	Dialog	$MR_{TotalCL}$	0.000	0.442	0.883	4.150
		$MR_{TotalFL}$	0.000	0.317	1.633	5.500
	QA	$MR_{TotalCL}$	0.092	0.267	0.783	2.525
		$MR_{TotalFL}$	0.100	0.133	1.250	3.092
	Classification	$MR_{TotalCL}$	0.000	0.000	0.000	0.000
		$MR_{TotalFL}$	0.000	0.000	0.000	0.000

Table 13: Memorization with various decoding methods(%)

Model	Decoding	Summarization		Dialog		QA		Classification	
		MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}
Qwen2.5-3B	temperature	0.475	0.067	1.267	1.442	1.283	0.750	0.000	0.000
	top-k	0.342	0.046	1.533	1.446	1.450	0.813	0.000	0.000
	top-p	0.525	0.050	3.792	2.996	2.567	1.438	0.000	0.000
Llama-3.2-3B	temperature	0.442	0.117	1.400	1.096	1.458	0.925	0.000	0.000
	top-k	0.700	0.067	3.167	2.313	2.183	1.375	0.000	0.000
	top-p	0.583	0.075	3.133	2.525	2.375	1.396	0.000	0.000

Table 14: Memorization with various prefix lengths(%)

Model	Prefix Length	Summarization		Dialog		Abstractive QA		Classification	
		MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}
Qwen2.5-3B	10	0.508	0.188	2.108	1.992	1.525	1.383	0.000	0.000
	30	0.342	0.046	1.533	1.446	1.450	0.813	0.000	0.000
	50	0.425	0.038	1.575	1.242	1.242	0.550	-	-
	100	0.208	0.004	1.408	1.150	1.125	0.429	-	-
Llama-3.2-3B	10	0.883	0.379	3.650	3.188	4.383	3.679	0.000	0.000
	30	0.700	0.067	3.167	2.313	2.183	1.375	0.000	0.000
	50	0.692	0.063	2.533	1.542	2.150	0.938	-	-
	100	0.275	0.017	3.167	2.342	2.300	0.567	-	-

Table 15: Memorization with different federated algorithm(%)

Model	Tasks	Federated Algorithm	MR_{Intra}	MR_{Inter}	Performance
Qwen2.5-3B	Summarization	FedAvg	0.342	0.046	29.88
		FedProx	0.942	0.138	30.51
	Dialog	FedAvg	1.533	1.446	18.11
		FedProx	1.892	1.879	17.47
	Abstractive QA	FedAvg	1.450	0.813	28.60
		FedProx	3.675	2.146	28.36
	Classification	FedAvg	0.000	0.000	51.22
		FedProx	0.011	0.000	80.16
Llama-3.2-3B	Summarization	FedAvg	0.700	0.067	28.67
		FedProx	2.442	0.413	30.78
	Dialog	FedAvg	3.167	2.313	19.44
		FedProx	8.900	6.979	18.91
	Abstractive QA	FedAvg	2.183	1.375	28.53
		FedProx	14.092	7.892	28.76
	Classification	FedAvg	0.000	0.000	78.16
		FedProx	0.300	0.117	83.14

Table 16: Intra- vs inter-client memorization in FL(%)

Model	Summarization		Dialog		QA		Classification	
	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}	MR_{Intra}	MR_{Inter}
Qwen2.5-0.5B	0.550	0.167	1.817	1.483	1.067	0.621	0.000	0.000
Qwen2.5-1.5B	0.992	0.113	1.817	1.521	1.225	0.717	0.000	0.000
Qwen2.5-3B	0.342	0.046	1.533	1.446	1.450	0.813	0.000	0.000
Llama-3.2-1B	0.800	0.088	3.808	2.938	2.108	1.275	0.000	0.000
Llama-3.2-3B	0.700	0.067	3.167	2.313	2.183	1.375	0.000	0.000

Table 17: Impact of communication rounds(%)

Model	Tasks	Communication Rounds	MR_{Intra}	MR_{Inter}	Performance
Qwen2.5-3B	Summarization	1	0.433	0.088	29.31
		3	0.342	0.046	29.88
		5	0.467	0.058	30.01
	Dialog	1	1.833	1.521	18.26
		3	1.533	1.446	18.11
		5	1.633	1.363	17.14
	Abstractive QA	1	1.300	0.650	28.68
		3	1.450	0.813	28.60
		5	1.108	0.763	28.37
	Classification	1	0.000	0.000	19.64
		3	0.000	0.000	51.22
		5	0.000	0.000	57.57
Llama-3.2-3B	Summarization	1	0.650	0.096	27.38
		3	0.700	0.067	28.67
		5	0.600	0.075	28.57
	Dialog	1	4.450	3.588	19.20
		3	3.167	2.313	19.44
		5	2.333	1.996	19.39
	Abstractive QA	1	2.542	1.550	28.37
		3	2.183	1.375	28.53
		5	1.983	1.263	28.34
	Classification	1	0.000	0.000	37.34
		3	0.000	0.000	78.16
		5	0.000	0.000	79.73

Table 18: Trade-off between centralized learning and FedAvg(%)

Model	Task	Framework	Bleu	RougeL	Memorization
Qwen2.5-3B	Summarization	$MR_{TotalCL}$	37.04	28.46	0.558
		$MR_{TotalFL}$	37.20	29.88	0.433
	Dialog	$MR_{TotalCL}$	30.55	19.40	3.417
		$MR_{TotalFL}$	28.87	18.11	3.992
	QA	$MR_{TotalCL}$	32.47	26.66	2.150
		$MR_{TotalFL}$	31.25	28.60	2.917
			Accuracy		
	Classification	$MR_{TotalCL}$	76.30		0.000
		$MR_{TotalFL}$	51.22		0.000
Llama-3.2-3B	Summarization	$MR_{TotalCL}$	35.03	26.80	0.942
		$MR_{TotalFL}$	36.95	28.67	0.808
	Dialog	$MR_{TotalCL}$	31.05	19.62	5.308
		$MR_{TotalFL}$	29.77	19.44	7.050
	QA	$MR_{TotalCL}$	32.80	26.67	3.458
		$MR_{TotalFL}$	32.94	28.53	4.300
			Accuracy		
	Classification	$MR_{TotalCL}$	77.83		0.000
		$MR_{TotalFL}$	78.16		0.000