CACHE ME IF YOU CAN: THE CASE FOR RETRIEVAL AUGMENTATION IN FEDERATED LEARNING

Aashiq Muhamed, Pratiksha Thaker, Mona Diab, Virginia Smith
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{amuhamed, pthaker, mdiab, smithv}@andrew.cmu.edu

ABSTRACT

We propose retrieval augmentation (RA) as an enhancement to federated learning (FL) that can improve privacy protection and ensure regulatory compliance. FL, primarily designed for data privacy preservation, faces challenges with conventional parametric models which are susceptible to privacy breaches and potentially non-compliant with regulations such as data erasure mandates. RA can help to address these issues by integrating a retrieval-based method during the inference phase, achieving "perfect secrecy" by limiting server access to private documents and reducing barriers to compliance. This study conducts a thorough evaluation of RA's efficacy within the FL paradigm, positioning it as a preferable alternative to traditional parametric models within analogous memory constraints. We characterize potential applications that may benefit from RA in FL, showing in particular that it is well-suited for knowledge-intensive, few-shot environments—offering scalable inference-time operations, source attribution, and the ability to dynamically update and unlearn knowledge for compliance. We present a new modeling framework, named *Raffle*, to investigate RA for FL applications with labeled and unlabeled data. Implementing *Raffle* in homogeneous settings for few-shot question answering, we explore the influence on client participation dynamics and the importance of passage index composition for effective generalization.

1 INTRODUCTION

Modern federated learning (FL) systems are designed to prioritize privacy, aiming to align closely with regulatory standards while maintaining performance (McMahan et al., 2016). FL incorporates concepts of data minimization and anonymization to ensure only essential data is aggregated for specific computations. In these systems, parametric models are preferred for their efficiency in transmitting necessary updates rather than the entirety of raw data. However, despite their widespread adoption, FL systems using parametric models face notable challenges. Although frameworks that communicate parametric models/model updates reduce data exposure, they are are prone to privacy breaches through membership inference (Shokri et al., 2016), model inversion (Jia & Gong, 2018), and attribute inference attacks (Fredrikson et al., 2015). Operating under stringent data governance frameworks like GDPR, FL must adapt to requirements such as the 'obligation of data erasure', which mandates the selective deletion of data upon user request, which is a complex task for parametric models (Ginart et al., 2019; Li et al., 2021; Bourtoule et al., 2021). Additionally, knowledge-intensive tasks like question answering often require massive parametric models which are expensive to host and unstable to finetune in low-data regimes, a scenario common in FL settings (Dodge et al., 2020; Mosbach et al., 2021).

To mitigate these privacy issues, there has been increasing research interest in using public datasets to augment learning on private data (Wang et al., 2023). A potential benefit of using public data is that, with enough public data, clients can avoid sharing private information entirely—using a model trained on public data and then augmenting it locally with private data. Retrieval Augmentation (RA) (Lewis et al., 2020) exemplifies this strategy by merging parametric and non-parametric memory systems to address the limitations of parametric models in FL. By leveraging public datasets for training and utilizing information retrieval and augmentation with private datasets at inference, RA

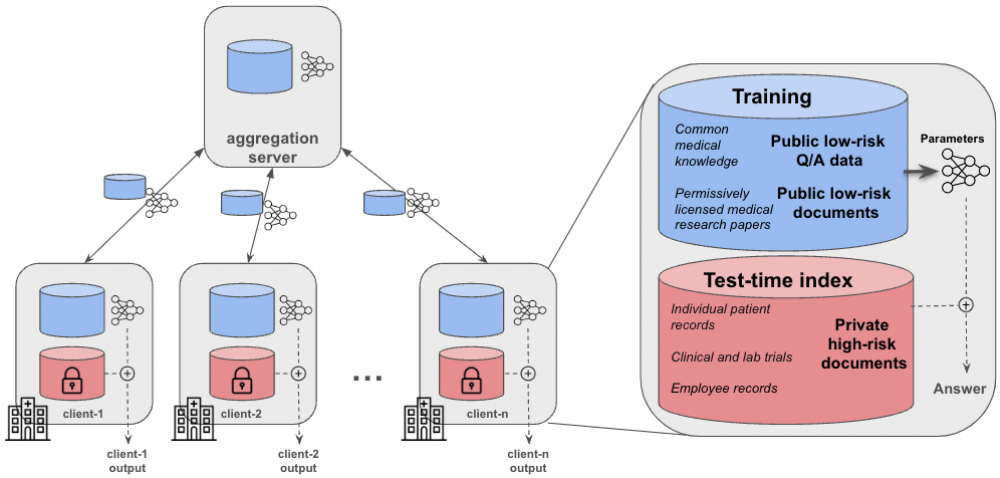

Figure 1: Raffle trains a parametric language model on client public datasets that contain non-sensitive information, which is then communicated via FL. At test time, a nonparametric index is used that can include private high-risk data. Dashed lines indicate inference time, and solid lines indicate training time.

not only enhances privacy but also ensures dynamic compliance. It can provide *perfect secrecy* by not revealing any additional information about private documents beyond the retrieved portions, thus offering strong protection against privacy attacks like membership inference. RA's capability to control information flow at inference time aligns well with GDPR requirements like selective data erasure and user access modifications. Additionally, RA excels in knowledge-intensive tasks with significantly fewer parameters, leading to efficient communication and superior performance, especially in low-data scenarios prevalent in FL (Izacard et al., 2022).

In this work, building upon RA research (Min et al., 2023; Wutschitz et al., 2023) for privacy in local language modeling, we propose *Raffle*, a framework for privacy and RA in FL (see Figure 1). *Raffle* allows clients to contribute both their supervised training sets and unsupervised passage data, and we apply it to few-shot open-domain question answering (QA). In open-domain QA, the retriever must first identify relevant passages from a broad corpus, and then construct answers from these passages, a method exemplified in hospital cross-silo QA systems governed by stringent privacy regulations. We elaborate on the various design choices and distribution assumptions in Raffle, and experiment on the homogeneous client distribution setting. Our results first establish that RA significantly outperforms traditional parametric client models on few-shot QA. Furthermore, we explore the impact of passage index composition beyond mere scaling, as previously examined, introducing a new dimension to FL. Our analysis reveals that the inclusion of relevant passages at training time improves generalization, whereas the incorporation of hard negatives can impair it. These insights prompt a reevaluation of traditional client participation incentives in FL, suggesting new directions for future research. Our contributions include:

- We advocate for RA in FL as an attractive alternative to standard parametric FL approaches within the same memory budget. This approach is especially effective for knowledge-intensive and few-shot applications, where it can achieve high utility and allow for inference-time scalability. RA also enables the ability to update and unlearn knowledge for compliance purposes, and incorporates built-in source attribution, all while ensuring perfect secrecy.
- We propose a modeling framework, *Raffle*, specifically designed to explore RA in FL across applications with both unlabeled and labeled data. This framework sheds light on various design decisions, each presenting potential avenues for further research.
- Utilizing a specific implementation of *Raffle* using a homogeneous benchmark, we identify applications we expect to benefit most from RA, such as few-shot, knowledge-intensive question answering. We evaluate the efficacy of our approach, including the impact of passage index composition and the dynamics of the federated incentive structure.

2 Background & Motivation

Retrieval Augmented Generation (RA) models (Lewis et al., 2020; Izacard et al., 2022) integrate a memory module within parametric LMs, facilitating the retrieval and incorporation of external

information to enhance the capabilities of the primary model. Our focus is on RA models where the parametric memory consists of a pre-trained seq2seq transformer, and the non-parametric memory comprises a dense vector index accessed via a pre-trained neural retriever. These components are trained end-to-end within a probabilistic model.

The sequence level RA model p_{RA} uses the retrieved document to generate the complete sequence. It treats the retrieved document as a single latent variable z that is marginalized to get the seq2seq probability $p(y|x)$ via a top-K approximation. The top K documents are retrieved using the retriever p_η, and the reader p_θ produces the output sequence probability for each document, which are then marginalized,

$$p_{\text{RA}}(y|x) \approx \sum_{z \in \text{top-}k(p(z|x))} p_\eta(z|x) p_\theta(y|z,x) = \sum_{z \in \text{top-}k(p(z|x))} p_\eta(z|x) \prod_{i=1}^{N} p_\theta(y_i|z,x,y_{1:i-1})$$

2.1 How can Retrieval Augmentation Help Federated Learning?

FL allows multiple participants to collaboratively construct a shared model without the need to centralize sensitive data. By empowering individual devices or entities to train models locally and share only model updates with a central aggregator, FL aims to use the ideas of data minimization and anonymization to prioritize privacy and align closely with regulatory standards while maintaining performance (McMahan et al., 2016).

Privacy. FL systems communicating model updates are vulnerable to privacy attacks (Shokri et al., 2016; Jia & Gong, 2018; Fredrikson et al., 2015; Chase et al., 2021; Carlini et al., 2018). Strategies to enhance privacy in FL include introducing statistical noise (Dwork et al., 2006) and utilizing public data to improve private training efficiency without compromising private information (Gu et al., 2023; Liu et al., 2021; Pinto et al., 2023; Ganesh et al., 2023; Wang et al., 2023). Utilizing public data can potentially eliminate the need for sharing private data by locally augmenting a publicly trained model with private information at inference. This approach, exemplified by RA ensures *perfect secrecy* by achieving user-level Differential Privacy with $(\varepsilon = 0, \delta = 0)$ (Tschantz et al., 2017).

Regulatory Compliance. Models must ensure privacy and comply with data governance laws like GDPR (EU, 2014), which mandates selective or complete data deletion based on governance policies. Traditional parametric models face challenges in implementing access control and adapting to dynamic access modifications, where user rights change over time. These challenges hinder scalability in environments with expanding user bases, as they often require frequent retraining or managing multiple models. RA offers a solution by enabling explicit access control policies during inference to generate compliant responses for multiple users (See Figure 2). This is achieved by pre-training the base retriever and reader on a permissive, low-risk public dataset D. Private datasets (denoted as N datasets D_i, each governed by a compliance policy P_i) are queried only during inference. This approach allows the model to surface information exclusively within the user's authorized scope, generating a P-compliant response O based on k retrieved samples for a given query Q.

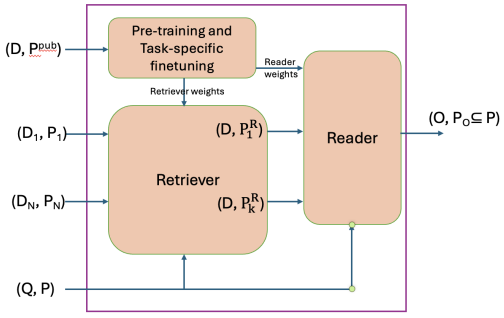

Figure 2: Information flow diagram: Public data D with policy P^{pub} used for training; N private datasets D_i with policies P_i, retrieved at test time.

Utility. Parametric FL faces a well-documented challenge: fine-tuning on limited private data leads to model instability, particularly in knowledge-intensive tasks requiring large parameter counts (Dodge et al., 2020; Mosbach et al., 2021). RA models excel in such tasks with significantly fewer parameters, demonstrating efficient few-shot generalization and matching the performance of larger models with up to 50x fewer parameters (Izacard et al., 2022). This translates to improved performance and data efficiency in FL, facilitating personalization to unique client data profiles (Salemi et al., 2023). Furthermore, RA models provide precise document source attribution during inference.

Algorithm 1 Raffle: Retrieval Augmented Federated Learning

Require: Shared pretraining data D_{shared}, Public dataset per client $\{D_{\text{public},i}\}_{i=1}^{M}$, Private dataset $\{D_{\text{private},ij}\}_{i=1,j=1}^{N,M}$ with policy $\{P_j\}_{j=1}^{N}$, Query Q, Retrieval number k
Ensure: P_j-compliant response O

Pretraining:
Pretrain retriever and reader on D_{shared} to learn a generalized representation.

Federated (or Local) Learning (Client-Specific Public Data Training):
for each client i with public dataset $D_{\text{public},i}$ **do**
  Perform federated (or local) finetuning to train parametric models on $D_{\text{public},i}$
end for

Inference (Test Time) with Private Data:
for each client i based on policy P_j **do**
  Identify user's authorized datasets $\{D_{\text{private},ij}\}$ based on P_j.
  Retrieve top k samples $\{S_l\}_{l=1}^{k}$ from $\{D_{\text{public},i}\} \cup \{D_{\text{private},ij}\}$ relevant to Q.
  Generate response O by integrating information from $\{S_l\}_{l=1}^{k}$ into reader
  Ensure O is P_j-compliant by filtering information based on access controls.
end for
return O

3 RAFFLE: RETRIEVAL AUGMENTED FEDERATED LEARNING

Raffle is a specific implementation of RA in FL (see Algorithm 1). It follows the text-to-text framework (Raffel et al., 2020), where the system receives an input text query and generates a text output. For instance, in question answering, the query corresponds to the question, and the model generates the answer. Raffle comprises a retriever and a reader. For QA, the model first retrieves the top-k relevant documents from a large corpus of text passages using the retriever, then feeds them, along with the query, to the reader, which in turn generates the output. Both the retriever and the reader are based on pre-trained transformers.

Retriever: Raffle employs the Contriever (Izacard et al., 2021), a dual-encoder model where the query and documents are independently embedded by a transformer encoder (Karpukhin et al., 2020). Average pooling is applied to the outputs of the last layer to obtain one vector representation. A similarity score between the query and each document is calculated by computing the dot product between their corresponding embeddings.

Reader: For the reader, Raffle utilizes the Fusion-in-Decoder modification (Izacard & Grave, 2020) of T5 models, processing each document independently in the encoder. The query is concatenated with each document, and cross-attention is performed over the encoder outputs corresponding to the different documents.

Raffle is pre-trained using a Perplexity Distillation loss (Izacard et al., 2022) for the retriever and a masked language modeling loss for the joint retriever and language model. It is pre-trained on 350 million passages from the 2021 Wikipedia dump and a subset of the 2020 Common Crawl dump (Thurner et al., 2018).

3.1 RAFFLE HOMOGENEOUS BENCHMARK

We evaluate Raffle on supervised fine-tuning (SFT) of RA models within an FL setting and propose a benchmark for question answering (QA) that incorporates both supervised QA pairs and unsupervised text passages. SFT is essential to enhance the few-shot generalization capabilities for smaller models used in FL. The benchmark is designed around distinct data modalities essential for task-specific fine-tuning: **Question Space** (X), consisting of natural language queries; **Passage Corpus** (Z), a collection of text passages forming the knowledge base; and **Answer Space** (Y), containing potential answers derived from Z. The data-generating process, denoted as $\mathcal{D}_{XYZ}$, involves sampling passages $Z \sim \mathcal{D}_Z$ and identifying corresponding question-answer pairs (x, y), where $x \sim \mathcal{D}_{X|Z}$ and $y \sim \mathcal{D}_{Y|X,Z}$.

To address privacy concerns in RA modeling, we incorporate two data compliance policies per client: Public, permissive, and non-restrictive, and Private, which is restrictive. The data (X), (Y), and (Z) are partitioned into public and private segments. The Homogeneous benchmark, is based on web queries with public and private segments that are disjoint yet identically distributed as $\mathcal{D}_{XYZ}$ across clients. For training via empirical risk minimization, public and private segments are further divided into training and testing sets, as depicted in Figure 3. Organizing data into four splits (A, B, C, D) per client, corresponding to quadrants in Figure 3, ensures unique training and testing examples for each split, along with the corresponding ground truth reference passages for each QA pair. The remaining passages are distributed proportionally across splits, and can be used to study the impact of passage composition.

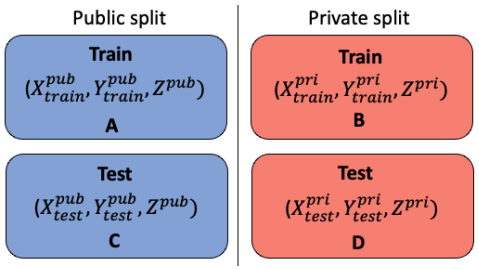

Figure 3: The four splits in the benchmark.

Threat model: We consider an FL setting with semi-honest adversaries where clients hold both public and private data. Public data (quadrant A) is used for training the model's parameters, while private data (quadrant D) is accessed only during inference. We view each client as a data silo containing data pertaining to multiple users. Privacy may be considered either at the example-level (e.g., protecting individual patient records) or the silo-level (e.g., preventing information sharing between hospitals). Public and private passage indices are merged during testing ($Z^{pub} \cup Z^{pri}$), with datasets indexed per client.

4 Experiments And Results

4.1 Datasets, Metrics, and, Hyperparameters

Few shot Q/A data We construct our datasets from the NaturalQuestions dataset (Kwiatkowski et al., 2019). Each client contains 64 train Q/A pairs, and we split the dev set (8752 pairs) and test set (3600 pairs) evenly among the 8 clients.
Passage data We use the Wikipedia 32M passages *wiki-dec2018* used in Izacard et al. (2022), split into public and private passages for specific experiments, and exact maximum inner product search to retrieve documents.
Metrics We employ two widely recognized metrics in QA systems: Exact Match and F1 Score. In both local and FL settings, we report the average of these metrics for the dev or test set across all clients. In the multitask setting, we use macro-averaging. For the local setting, we average the best scores achieved by each client. In the FL setting, we report the metrics from the round that yields the highest average Exact Match score.
Training Both the parametric and RA models use similar hyperparameters, employing AdamW with a batch size 64 and an lr of 4×10^{-5} with cosine decay. The local models are trained for 1000 steps, while the FL models are trained for 10 rounds with 64 steps/round. For Raffle few-shot finetuning we also train the query encoder of the retriever and retrieve 40 nearest neighbors passages from the index for every question. The local models are evaluated every 100 steps, while the federated models are evaluated every round. We use FedAvg (McMahan et al., 2016) at the server and train on 8 clients. Full training details can be found in the Appendix.

4.2 Few-shot learning

In Table 1, we compare the few-shot performance (64-shot) of Raffle against parametric models without a private/public split. We consider models of about the same size (about 220M parameters). Raffle retrieves relevant documents from the entire Wikipedia index using a dense retriever. We additionally compare against t5-lm-adapt-base (Raffel et al., 2020), which was more stable to fine-tuning than t5-base, and flan-t5-base (Chung et al., 2022), a strong instruction tuned model, both of which are closed-book. The evaluation is structured across three training configurations: Centralized (combining datasets from all clients for unified model training and evaluation), Local (training and evaluating models on individual client datasets), and Federated (where models are locally trained, aggregated after each round, and assessed on local test sets).

Model name	Centralized		Local		Federated	
	Exact match	F1	Exact match	F1	Exact match	F1
t5-base	2.361	5.892	1.694	4.631	2.639	6.599
flan-t5-base	3.250	7.478	3.361	7.487	3.917	8.497
Raffle	32.556	41.071	28.639	36.178	31.639	39.900

Table 1: Few-shot performance of retrieval-augmented Raffle and parametric models t5 and flan-t5. Raffle consistently outperforms parametric models. FL yields more substantial improvements for Raffle than for parametric models.

Overall, we find that Raffle consistently outperforms parametric-only baselines across training configurations. In knowledge-intensive Q/A tasks, parametric models typically need significantly more parameters to achieve comparable results (Lewis et al., 2020). Notably, flan-t5 exhibits only a marginal improvement over t5-base. Furthermore, we find that when using the entire Wikipedia index, there is an advantage to FL over local training alone, particularly when the number of labeled Q/A pairs per client is limited. This advantage is anticipated to be more pronounced under greater data scarcity. The performance of the FL baseline closely matches that of the Centralized baseline, a trend previously noted in scenarios with homogeneous client data. Performance metrics on the dev set are available in Appendix A.5.

4.3 Effect of Passage Index

Index	Public Index	Private Index
REL	Relevant + 80% Wiki (Shared)	Relevant + 2.5% Wiki (Unique)
IRR	80% Wiki (Shared)	Relevant + 2.5% Wiki (Unique)
REL-1	7 clients: 80% Wiki (Shared); 1 client: All Relevant + 80% Wiki	Relevant + 2.5% Wiki (Unique)
SPLIT	Relevant + 10% Wiki (Unique)	Relevant + 2.5% Wiki (Unique)

Table 2: Raffle client index options. Each index, consists of relevant passages for train or dev Q/A and a % of the rest of Wikipedia. The indexes can be unique to each client or shared across multiple clients.

To investigate the impact of passage composition on few-shot QA performance, we use the task formulation outlined in Section 4.2 and create four custom splits as detailed in Table 2, featuring disjoint public and private passages. Our focus is to understand how the presence of relevant and irrelevant passages, associated with the public train set, affects test set performance. Each index, assigned to every client, consists of train or dev Q/A nearest neighbor passages and a proportion of the rest of Wikipedia. The indexes can either be unique to each client or shared across multiple clients. We employed the BM25 algorithm on concatenated Q/A pairs to retrieve the most relevant nearest neighbor passages and verified that the top 5 neighbors contained the answer to the question. Should a public and private passage index contain identical passages during construction, we eliminate the duplicate from the public index. The composition of private passages remains constant across all four settings. The REL index serves as a baseline federated index. In IRR, all relevant passages from REL, of the train set, are excluded. REL-1 is a variant of IRR, where one client possesses all relevant train passages from every client. In SPLIT, the REL index is distributed across eight clients, each holding relevant passages specific to its train set.

In Table 3, we compare the performance of three Raffle baselines across various index options. Raffle Local is the averaged local model performance trained without federation, Raffle Local Merged uses the merged public and private passage index for training, Raffle Fed is trained using FL. We find that FL consistently improves performance in few-shot settings compared to local baselines. Merging private and public passages locally results in minimal improvement, as adding irrelevant passages during training predominantly increases noise. The performance of IRR, which excludes all relevant passages, significantly declines relative to REL which indicates that relevant passages corresponding to the train Q/A dataset are critical for the model's learning and generalization capabilities. In the REL-1 configuration, where one client holds all relevant passages, there is only a

Index	Raffle Local		Raffle Local Merged		Raffle Fed	
	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match	F1 Score
REL	29.627	39.985	30.222	40.128	34.826	46.357
IRR	23.069	32.014	23.891	32.114	30.085	40.175
REL-1	24.802	33.634	24.394	32.376	31.627	41.224
SPLIT	33.355	43.344	33.021	42.779	39.145	49.570

Table 3: Average client dev performance with the four index options on Raffle Local, Raffle Local Merged, and Raffle Federated. Federation improves performance across index options. The presence of relevant passages in REL boosts performance over IRR. Using SPLIT index performs better than a federated REL index.

marginal improvement over IRR. This is because the improvement occurs through a longer information channel: relevant passages retrieved by client 8 improve generalization in client 8, which is then communicated to other clients through FL. Employing a direct channel for relevant passages, as seen in the federated REL index, substantially improves performance in comparison. Finally, SPLIT, with a higher ratio of relevant to hard negative passages, markedly improves performance. This suggests that ensuring the presence of relevant passages and pruning hard-negative passages can further enhance performance (Cuconasu et al., 2024).

Note on Incentives: The observed trends with passage composition alter the incentives for client participation in FL. Traditionally in FL, particularly in homogeneous settings, clients are motivated to participate to maximize shared train Q/A data. However, sharing passages modifies these incentives. A client possessing all relevant passages for their train set would find participation counterproductive, as it would introduce hard negatives into their index. Conversely, if a client has few relevant passages, participating in a federated index offers significant benefits. However, beyond a certain index size, marginal returns may be observed, contingent on the composition of hard negatives and irrelevant passages in the shared index. Generally, participation would depend on the number of relevant local public passages, the ratio of relevant to irrelevant local public passages, the volume of local training data, and the composition of the federated index. Future research should explore techniques to refine client passage indexes for increased relevance and performance and address performance poisoning through adversarial passage injection (Zhong et al., 2023).

5 Conclusion and Future Work

In this work, we propose RA as an effective alternative to parametric modeling in FL, particularly for knowledge-intensive and few-shot scenarios—offering benefits like scalability, compliance, and perfect secrecy. In future work, we would like to train and evaluate on additional federated retrieval augmented generative tasks beyond QA and explore other directions such as understanding the effect of data heterogeneity in public and private data; considering effective ways to train on private labeled data; and studying architectural decisions such as whether maintaining separate retrievers for the public and private passage index can further boost generalization.

References

Simran Arora and Christopher Ré. Can foundation models help us achieve perfect secrecy? *arXiv preprint arXiv: Arxiv-2205.13722*, 2022.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Nicholas Carlini, Chang Liu, Ú. Erlingsson, Jernej Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. *USENIX Security Symposium*, 2018.

Melissa Chase, Esha Ghosh, and Saeed Mahloujifar. Property inference from poisoning. *arXiv preprint arXiv: 2101.11073*, 2021.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv: 2210.11416*, 2022.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv: 2401.14887*, 2024.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv: 2002.06305*, 2020.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pp. 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540327312. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

EU. Council regulation (EU) no 269/2014, 2014. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502&uri=CELEX:32014R0269.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015. URL https://api.semanticscholar.org/CorpusID:207229839.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pp. 10611–10627. PMLR, 2023.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Xin Gu, Gautam Kamath, and Zhiwei Steven Wu. Choosing public datasets for private machine learning via gradient subspace distance. *arXiv preprint arXiv:2303.01256*, 2023.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *Conference of the European Chapter of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2021.eacl-main.74.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2021.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language model. *arXiv preprint arXiv: Arxiv-2208.03299*, 2022.

Jinyuan Jia and N. Gong. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. *USENIX Security Symposium*, 2018.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906, 2020. URL https://api.semanticscholar.org/CorpusID:215737187.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, and Steven Wu. Leveraging public data for practical private query release. In *International Conference on Machine Learning*, pp. 6968–6977. PMLR, 2021.

H. B. McMahan, Eider Moore, Daniel Ramage, S. Hampson, and B. A. Y. Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, 2016.

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. Silo language models: Isolating legal risk in a nonparametric datastore. *arXiv preprint arXiv: 2308.04430*, 2023.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines, 2021.

Francesco Pinto, Yaxi Hu, Fanny Yang, and Amartya Sanyal. Pillar: How to make semi-private learning more effective. *arXiv preprint arXiv:2306.03962*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. *arXiv preprint arXiv: 2304.11406*, 2023.

R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *IEEE Symposium on Security and Privacy*, 2016. doi: 10.1109/SP.2017.41.

Stefan Thurner, Rudolf Hanel, and Peter Klimekl. Scaling. *Oxford Scholarship Online*, 2018. URL https://api.semanticscholar.org/CorpusID:239790883.

Michael Carl Tschantz, Shayak Sen, and Anupam Datta. Differential privacy as a causal property. *arXiv preprint arXiv: 1710.05899*, 2017.

Boxin Wang, Yibo Jacky Zhang, Yuan Cao, Bo Li, H Brendan McMahan, Sewoong Oh, Zheng Xu, and Manzil Zaheer. Can public large language models help private cross-device federated learning? *arXiv preprint arXiv:2305.12132*, 2023.

Jason Wei, Yi Tay, and Quoc V. Le. Inverse scaling can become u-shaped. *Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.48550/arXiv.2211.02011.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzm'an, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *International Conference on Language Resources and Evaluation*, 2019.

Lukas Wutschitz, Boris Köpf, Andrew Paverd, Saravan Rajmohan, Ahmed Salem, Shruti Tople, Santiago Zanella-Béguelin, Menglin Xia, and Victor Rühle. Rethinking privacy in machine learning pipelines from an information flow control perspective. *arXiv preprint arXiv:2311.15792*, 2023.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. *Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.48550/arXiv.2310.19156.

A Appendix

A.1 Related Work

In-context learning (ICL) on locally hosted foundation models preserves data, task, and label privacy, and achieves perfect secrecy (Arora & Ré, 2022). ICL alone may not be sufficient for real-world tasks that often involve document passages and labeled data for fine-tuning the model, where participating in FL to share data could improve generalization. Hosting foundation models with billions of parameters is resource-intensive; training them with few-shot data is often unstable, fine-tuning them on passages can be prohibitively expensive, and editing or unlearning knowledge in them remains an open problem. Recent RA models (Izacard et al., 2022) have been shown to outperform foundation models that are 50 times larger when access to an external passage corpus is available.

Privacy preserving RA Silo-LM (Min et al., 2023) trains a parametric language model (LM) on low-risk data and incorporates high-risk data only at inference through a nonparametric component. They show that datastore size predictably reduces LM perplexity. Wutschitz et al. (2023) examines privacy from an information flow control perspective, and compares a zero-shot baseline and a full-finetuned baseline with an RA model for language modeling. They find that RA architectures when applied at inference time, offer superior utility and scalability while achieving perfect secrecy. In both works, the RA model is either local or a shared global model, which may not always be practical. Clients are less incentivized to share raw data, and using a shared model often implies hosting larger parametric models, potentially sacrificing personalization, especially under data or task heterogeneity. In our work, clients control their participation, both in terms of training on public Q/A data and passage data. Additionally, in our framework, clients train on supervised and unsupervised data, unlike the sole use of unsupervised data in language modeling. Our experiments focus on the QA task because a) LM perplexity does not always correlate with improved client task performance (Wei et al., 2022), and b) scaling the datastore can hurt task performance.

A.2 Training Details

For question answering, we format the input using the following template:

```
question: {question text} answer: [MASK_0]
```

and train the model to generate the mask token followed by the answer:

```
[MASK_0] {answer}.
```

We generate answers using greedy decoding. For both training and testing, we retrieve 40 passages and truncate the result of the concatenation between the query and the passages to 384 tokens.

All models are trained with bf16 precision. For few-shot local fine-tuning, we train Raffle for 1000 steps using 64 random samples from the train sets. The retriever is trained using query-side fine-tuning. We use AdamW with a batch size of 64 and a learning rate of 4×10^{-5} with linear decay for both the language model and the retriever. The local models use 20 iterations of warmup, while there is no warmup in FL. The local models are trained for 1000 steps, while the FL models are trained for 10 rounds with 64 steps/round. All models can be trained using 4 A6000 in under a day.

A.3 PRETRAINING DATA

All models were trained on Common Crawl (Wenzek et al., 2019), which includes English Wikipedia. However, the models do not perform well on NaturalQuestions without finetuning, as the content and style of Wikipedia articles differ significantly from the types of queries found in the Natural Questions dataset. Additionally, the training objectives and data used for pretraining do not align perfectly with the requirements of answering natural questions.

A.4 SYSTEMS CONSIDERATIONS

Retrieval augmentation introduces additional memory requirements for hosting the query and document encoders, and the passage index. These can be managed by offloading to the CPU. We utilize exact MIPS search, but index compression techniques (e.g., Douze et al. (2024)) offer feasible alternatives for cross-device FL. RA models might increase inference time; however, employing Fast ANN methods, quantization, and pruning can effectively mitigate this, ensuring efficiency in computationally constrained environments.

A.5 FEW SHOT LEARNING

In this section we include the dev set metrics for few-shot QA model performance.

Model name	Centralized		Local		Federated	
	Exact match	F1	Exact match	F1	Exact match	F1
t5-base	1.862	4.986	1.302	3.814	2.057	5.343
flan-t5-base	3.142	7.069	2.959	6.852	3.736	7.956
Raffle	32.735	41.594	28.222	37.219	31.936	41.125

Table 4: Model Performance on Dev Set