CAN FOUNDATION MODELS HELP US ACHIEVE PERFECT SECRECY?

Simran Arora  
Stanford University  
Stanford, CA  
simran@cs.stanford.edu

Christopher Ré  
Stanford University  
Stanford, CA  
chrismre@cs.stanford.edu

May 30, 2022

ABSTRACT

A key promise of machine learning is the ability to assist users with personal tasks. Because the personal context required to make accurate predictions is often sensitive, we require systems that protect privacy. A gold standard privacy-preserving system will satisfy perfect secrecy, meaning that interactions with the system provably reveal no additional private information to adversaries. This guarantee should hold even as we perform multiple personal tasks over the same underlying data. However, privacy and quality appear to be in tension in existing systems for personal tasks. Neural models typically require lots of training to perform well, while individual users typically hold a limited scale of data, so the systems propose to learn from the aggregate data of multiple users. This violates perfect secrecy and instead, in the last few years, academics have defended these solutions using statistical notions of privacy — i.e., the probability of learning private information about a user should be reasonably low. Given the vulnerabilities of these solutions, we explore whether the strong perfect secrecy guarantee can be achieved using recent zero-to-few sample adaptation techniques enabled by foundation models. In response, we propose FOCUS, a framework for personal tasks. Evaluating on popular privacy benchmarks, we find the approach, satisfying perfect secrecy, competes with strong collaborative learning baselines on 6 of 7 tasks. We empirically analyze the proposal, highlighting the opportunities and limitations based on task types, and model inductive biases and sizes.

1 Introduction

A key promise of machine learning is the ability to assist users with personal tasks. Personal tasks span familiar tasks such as topic classification over personal messages and open-ended question answering using personal context, to idiosyncratic tasks for user-specific needs. Given the private nature of the personal data required for such tasks, these systems should satisfy three desiderata: (1) no leakage of private information, (2) quality, and (3) feasibility. Guided by the desiderata, we present FOCUS, a framework for privately serving personalized tasks using recent foundation models [8].

The ideal privacy system will offer perfect secrecy — as users interact with the system, the probability that adversaries learn private information does not increase — and universal compositionality — as multiple personal tasks are completed using the same underlying private data, the probability of leakage does not increase [64][12]. A trivial way to achieve this classical privacy guarantee is through purely training or fine-tuning a model on a user’s private dataset. However, recent neural models require copious amounts of training data [17] and users often hold a limited scale of labeled data. [1]

To address the challenge that individual users lack sufficient data, federated learning (FL) over data spanning multiple privacy scopes (i.e. users) has emerged as a popular framework [65][59]. Requiring all users to ship data to a central location sacrifices privacy, so instead, FL trains a task model by shipping the model between users and a central server.

1On average across our evaluation tasks, which are popular benchmarks in the privacy and federated learning literature, an individual user has 149 (standard deviation 191) labeled training examples (See Table 4).
FOCUS Framework: logical privacy

Public Foundation Model

Unidirectional Information Flow

Users owning Private Data

Describe Tasks in Natural Language

Serve Multiple Personal Tasks

Federated Learning: statistical privacy

Shared Model on Central Server

Users Send Model Updates Computed on Private Data to the Central Server

Users Share a Task Goal

Unique Model per Personal Task

In Comparison: degrees of protection

“Inputs” Data privacy

“What” Task privacy

“How” Label privacy

Intuitive privacy parameters

Figure 1: Towards the goal of privately serving personal tasks, FOCUS proposes a unidirectional data flow architecture that entails shipping FMs to users and guarantees perfect secrecy.

In each training iteration, the model updates that are computed over each user’s data are aggregated at the central server. Along the desiderata:

• Privacy While FL prevents raw data from being shipped between devices, it does forgo perfect secrecy. Unfortunately, private information can be recovered from the exposed model [66, 51, 55, 6]. Formally, instead of using classical logical notions of privacy [53], significant effort is devoted to reasoning about these procedures using statistical notions of privacy [18]. These methods demonstrate a performance-quality tradeoff and the world’s largest organizations struggle to reason about how to set the appropriate privacy parameters when implementing them [25, 50].

• Quality FL improves model performance for the average user. However private data often widely differs by individual and performance is often quite uneven across participants [80]. Further, adversarial participants and central servers can corrupt the training procedure [75].

• Feasibility Considering systems costs, FL requires many rounds of communication between many users to perform well [33], introducing standard distributed systems challenges such as device heterogeneity and synchronization. These costs are incurred for every personal task a user wishes to perform. Considering incentives, FL requires aligning users with similar task goals and may exclude users who have no training data to contribute [78].

Noting these challenges, this work revisits the questions of when and where perfect secrecy is achievable, if at all. We observe natural compatibilities between the capabilities of foundation models (FM) and the above challenges. FMs are large models trained on massive amounts of data in a self-supervised manner and provide impressive performance on new downstream tasks — often out-of-the-box or with minimal effort [10, 58, 79, 44]. Recent FMs can perform a task given a textual description of the task (e.g., “Summarize my email.”) and zero-to-few demonstrations of the the task, with no gradient updates to the FM weights. Interestingly, these techniques have proven effective in tasks widely different from the task for which the FM was trained [14, 45].

Our goal is to understand if FMs can be applied to achieve strong privacy guarantees for personal tasks. We study the behavior of these nascent models in the standard FL setting using popular off-the-shelf FMs [56, 72] and methods for FM adaptation [10, 29, 48, 58]. We propose Foundation model Controls for User Secrecy (FOCUS), a framework for privately serving personal tasks, that consists of a unidirectional data flow architecture. FOCUS entails shipping off-the-shelf public FMs to private user silos and applying zero-to-few sample FM adaptation techniques, with the zero-to-few training examples users have available, to perform personal tasks. The framework is technically very simple — the meat of our investigation is whether it satisfies the desiderata:

• Privacy: What privacy guarantees does FOCUS provide? We formalize the privacy guarantee by employing the classical Bell-LaPadula model [7], which provides perfect secrecy. BLP is a simple and efficient model (details in Section 3), originally developed for government agencies to manage multi-security-level access control, mapping to our setting of publicly accessible FMs and privately accessible personal data. Beyond data privacy, the FM-baseline hides what a user is doing (task privacy) and how they are doing it (label schema privacy). Importantly, in contrast to FL, FOCUS does not require practitioners to set any privacy parameters.
We identify compatibilities between recent FM capabilities and standard privacy challenges and we define the F

2 Background and Related Work

Our focus is on applying public FMs to private tasks. Other work studies the challenge of pretraining FMs over sensitive data [41]. That question is beyond the scope of our paper.

Perfect secrecy A system preserves perfect secrecy if the probability of an adversary obtaining knowledge of a client’s private data does not increase as the client interacts with the system [64]. As the system performs multiple tasks over the same underlying private data, the probability an adversary obtains knowledge of the private data should not increase [12]. Mathematically, these are logical notions of privacy [53]. The Bell-LaPadula (BLP) model we employ in this work (detailed in Section 3) is an access-control based privacy framework that satisfies the above privacy properties [7].

Machine learning for personal tasks The above properties are achievable through purely training or fine-tuning on only the user’s private data, for which existing work assumes users own a sufficient amount of private labeled training data [73]. Fine-tuning can be unstable in the low-data regime [17, 54], the focus of our work. More importantly, storing a unique task-specific model for every personal task the user requires is expensive. Federated learning (FL), the popular approach for personal tasks in our setting, trains a model over the private data owned by multiple parties [49]. FL proceeds by initializing a global model \( \theta \) and solving the global objective:

\[
\min_{\theta} G(L(\theta; D_1), ..., L(\theta; D_i), ..., L(\theta; D_N))
\]

where \( L(\theta; D_i) \) are the local objectives for each client, and \( G(\cdot) \) is a function of the local objectives. During training, each client receives a copy of model \( \theta_i \) at the current timestep \( t \) from a central server, performs gradient descent on their local data \( D_i \), sends the gradients the central server. The central server uses \( G(\cdot) \) to combine the local gradients to produce \( \theta_{t+1} \). FedAvg [49], a weighted sum of the local objectives, is the vanilla choice for \( G(\cdot) \). Li et al. [40] describes alternate objectives.

To achieve personalization in FL, popular approaches include using a mixture of local and global model updates, training a global model with model-agnostic meta-learning so it may be quickly adapted to local data, and initializing the global model with a pretrained initial model [20, 80, 28]. These methods improve upon training from scratch with FedAvg, though still expose computations over user data and require storing a unique model per personal task. Empirically, the improvements of these methods over FedAvg also appear to degrade in increasingly non-IID settings [13, 3].

The key constraint under FL is that the raw data \( D_i \) never leaves the private silo for client \( i \), however, functions of \( D_i \) may be exposed. Unfortunately, attacks on the gradients or inference attacks on the final model output can the reveal private information [60, 51]. Techniques to protect the privacy of the communicated parameters include differential privacy (DP) [13], which degrades model performance [50], or secure multiparty computation [9], which does not protect against the membership inference attacks and significantly increases the computational cost [21].

Foundation Models We use popular existing FMs and FM adaptation techniques in our evaluations. In particular, we use simple zero-shot [58], in-context learning [10] and lightweight tuning [29, 48] methods (defined in Section 3). We are inspired by work on versatile FM systems with natural language interfaces [37, 77], though instead of ML methods, our focus is the consequences of these FM capabilities on privacy. We observe the need for rigorously benchmarking FMs on personal tasks.

https://github.com/simran-arora/privacy_fm
3  **FOCUS Framework**

In this section, we describe FOCUS, including the assumptions, architecture, and privacy guarantee.

### 3.1 Setup

**Objective** We focus on supporting personalized tasks and differentiate this from the problem of releasing a public model that has been trained on sensitive data. An ideal solution should provide privacy, quality along machine learning metrics, and be practical from a systems perspective. Consider client $i$ who wants to complete a suite of personalized tasks $t \in T_i$. For $t$, the client has a private dataset $D_{it} = \{x_j, y_j\}_{j=1}^{n_i}$, where $n_i \in [0, k]$ for some small $k$. Canonical tasks include intent or content classification, message completion, and question-answering. FOCUS is competitive in settings with the following properties:

1. **Distributions** The client task and data distributions can be independently chosen. This differs from FL, which generally requires users to share similar task goals and task label schema.

2. **Data scale** A client’s training dataset size, $n_i$, is small enough that training or fine-tuning purely locally results in low-quality models [17, 54].

3. **Privacy** The user’s data and/or tasks are privacy sensitive. In our trust model, any central party that mediates between the decentralized clients, other clients, and eavesdroppers are untrustworthy.

4. **User resources** Users have sufficient compute and memory resources to privately use the FM. We study the systems feasibility in detail in Section 4.1.

### 3.2 Architecture

The FOCUS architecture centers on a unidirectional dataflow, as depicted in Figure 1. FOCUS downloads the FM to the user’s private silo and uses simple zero-to-few sample FM adaptation techniques to perform the task, using the provided descriptions. Users then receive personal reasoning by expressing task descriptions or task demonstrations, through a natural language interface.

**Proof of concept** There are many possible technical implementations within the proposed setup. Given the lack of prior FM baselines to FL, we aim to compare vanilla baselines for both paradigms. In this work, our proof of concept uses the following popular existing FM adaptation methods:

* **Similarity search** (zero shot) This method uses the FM to generate embeddings of each class description and of inputs at inference time, computes the similarity between these embeddings, and outputs the class description that is most similar to the input embedding.

* **Prompting** (zero-to-few shot) This method conditions a language FM to generate textual answers for downstream tasks. Two basic approaches are: (1) provide the textual task instructions, e.g., “Classify my email”, (i.e. zero labeled examples), or (2) provide sequence of a few task-demonstrations. The latter shows the model the desired label format and properties of the data-distribution.

* **Lightweight-tuning** (few shot) Inference-only strategies are temperamental. With enough training examples, we can train a model for our task to avoid the handcrafting. Several recent works propose to fine-tune a small number of parameters (a subset of the existing FM or newly initialized) [29, 30, 38]. We specifically require lightweight tuning approaches that suffice with few-samples 48, 43.

### 3.3 Privacy Model

FOCUS’s privacy guarantee is extremely simple and intuitive from user and legal perspectives — no private data leaves the user device, guaranteeing perfect secrecy. We employ the classical Bell-LaPadula Model (BLP) access-control framework, which preserves perfect secrecy and universal compositionality [7]. BLP manages the access of subjects with assigned clearance levels to objects of assigned security levels and is defined by three security rules: (1) **Simple Security Property** subjects cannot read data at higher security levels, (2) ***-Property** subjects cannot write to data-stores at lower security levels, and (3) **Discretionary Security Property** discretionary access to objects can be granted or revoked from subjects. The model, and access control models more generally, have been widely and successfully used in high-risk settings for decades [31].

**Analysis** In our setting, the FM is owned by public entities, and users own zero-to-few private task demonstrations and task descriptions, which remain on device. Respecting the Simple Security Property, no private data or function of private data is exposed to untrustworthy subjects. FL violates this rule by exposing functions of the private data to
untrustworthy entities. Respecting the *-Property, untrustworthy entities cannot write information to the user under our framework. FL violates this rule by allowing the central server and other users to tamper with the information incorporated in the user’s local model. Additionally task privacy (what the user is doing) and label schema privacy (how is the user doing it) immediately emerge from our setup; neither is protected under FL.

4 Experiments and Analysis

To study the benefits and limitations of FMs for privacy in FOCUS, we ask the following questions:

1. Is FOCUS competitive with FL? How does this vary by FM size and bias, and task properties?
2. To what degree does FOCUS enable personalization to the user’s data distribution?
3. To what degree does FOCUS enable personalization to the user’s task distribution?

4.1 Experimental Setup

**Benchmarks** We use a representative set of standard benchmarks in the privacy literature, which are useful proxies for personal tasks such as intent and content classification, message completion, and QA [11, 27, 35, 42]. Each benchmark contains a small number of examples per user with mean 149, standard deviation 191 examples across users and tasks. We use: **Text classification**: We use the Sentiment140 2-way [24] and 20News 20-way classification benchmarks [36]. **Image classification**: We use the CelebA binary [47], CIFAR10 10-way [34], and Federated EMNIST 62-way classification benchmarks [15]. **Language modeling**: output the next word in the input sequence; we use the Reddit benchmark [11]. **Reading Comprehension**: given a natural language question and document, output the answer span in the document; we use the MRQA benchmark [22].

**Foundation Models** We use a representative set of FMs including T0 (3B and 11B parameters) [63], Grover (124M and 1.5B parameters) [76], and GPT-3 (125M, 1.3B, 2.7B, 6.7B and 175B p.) [10] for prompting tasks, MPNet-base bi-encoders for textual zero-shot similarity-search (110M p.) [60, 67], and CLIP for image zero-shot similarity-search (150M p.) [58]. We run models either using 1 40-GB A100 GPU or the OpenAI API. To avoid dataset leakage, i.e. data observed during FM pretraining contaminating the distinction between public and private information, we use reported details about the pretraining corpora to appropriately select FMs for tasks. We release our code and prompts.

**Federated Learning Baselines** The FL baselines include modern pretrained model architectures. We take the best numbers that use vanilla FedAvg for the task. The vision tasks use ViT(S) (22M p.) and ResNet101 (45M p.), text classification tasks use DistillBERT for 20News and MRQA (67M p.), and Stacked-LSTM for Sent140 and next-word prediction (1M p.). Models in Table 1 are trained over 10k-1Ms of examples, 5-1000s clients and 10-1000s rounds of model communication across tasks.

An extended discussion of the baselines is in Appendix [7], and a concrete memory and computation comparison is in Section 5. We note the FL baselines do not use objectives beyond FedAvg, yet face no differential privacy, adversarial clients or central server, communication compression techniques, or for certain benchmarks, even heterogeneity across clients, all of which degrade performance [50, 39, 62, 68]. Our FM baselines similarly use no advanced adaptation techniques [79, 46, 37]. For the first comparison of these paradigms, we focus on popular vanilla baselines for each.

4.2 Is FOCUS competitive with federated learning?

We try the simplest zero-shot baselines by either providing a prompt to the model or by computing the similarities between the embedded input to the embedded class descriptions, as explained in Section 3. Results in Table 1 show that in the zero-shot setting, FM performance competes with FL performance on 6 of 7 benchmarks. Critically, the above FOCUS numbers provide perfect secrecy and the FL numbers provide no privacy. However, certain FMs fail on F-EMNIST and 20News.

**Analysis** A challenge for enabling practical systems in FOCUS is how to privately select an FM for personal use. The FMs we evaluate vary along two dimensions: (1) model size and (2) inductive bias, and we observe a tradeoff between

---

3We use GPT-Curie and GPT-Davinci from OpenAI API as GPT-6.7B and GPT-175B respectively [56].

4A particular challenge was obtaining an FM for Reddit, since most FMs are trained on generic Common Crawl — we thus use Grover FMs, which are pretrained only on news articles and find the performance of Grover models directly match or exceed the performance of GPT models of the same size — 9.6% for GPT-125M vs. 11.4% for Grover-124M, and 13.2% for GPT-1.3B vs. 13.4% for Grover-1.5B. Due to the limited range of Gover model sizes and confirming the comparable results of both models, we proceed with GPT variants.

5Referencing evaluation suites which have curated the results across many existing works [27, 11, 42, 35].
Benchmark | Foundation Models | Method | Accuracy | Federated Learning | Method | Accuracy
--- | --- | --- | --- | --- | --- | ---
Reddit | Promoting (175B) | 13.6 | FedAvg (1M) | 11 | 13.4 |
MRQA* | Promoting (175B) | 64.1 | FedAvg (67M) | 42 | 27.1 |
Sent140 | Promoting (175B) | 80.4 | FedAvg (1M) | 52 | 69.5 |
20News | Promoting (175B) | 31.3 | FedAvg (67M) | 42 | 51.4 |
Sent140 | Promoting (11B) | 78.2 | FedAvg (1M) | 52 | 69.5 |
20News | Promoting (3B) | 19.1 | FedAvg (67M) | 42 | 51.4 |
Sent140 | Similarity-search (110M) | 61.5 | FedAvg (1M) | 52 | 69.5 |
20News | Similarity-search (110M) | 63.4 | FedAvg (67M) | 42 | 51.4 |
CelebA | Similarity-search (150M) | 86.7 | FedAvg (22M) | 57 | 86.6 |
CIFAR10 | Similarity-search (150M) | 88.2 | FedAvg (22M) | 57 | 98.5/86.9** |
F-EMNIST | Similarity-search (150M) | 22.3 | FedAvg (45M) | 62 | 81.0 |

Table 1: Test results on standard privacy benchmarks with zero-shot FMs use. Red indicates an FM with relatively weak inductive bias (e.g. the next-word-prediction objective), and Blue indicates an FM with relatively strong inductive bias (e.g. tuned on instruction-label pairs). *F1 score. **Trained from scratch — The reported vision models are pretrained by default.

![Sent140 Sentiment Classification](image1.png)  ![20News Topic Classification](image2.png)

Figure 2: Zero shot performance by model size and by the strength of the FM inductive bias. Red indicates relatively weak (more flexible) and Blue relatively strong inductive bias FMs (less flexible).

using smaller FMs trained with stronger inductive biases and model flexibility. Along (1), large models reliably provide quality, yet are not always accessible to low-resource users (Figure 2). Along (2), Figure 2 shows FMs with relatively strong inductive biases in blue and weak in red, as determined by the pretraining scheme. For example, T0 is fine-tuned on pairs of task-descriptions and outputs and it is challenging to use these models outside this format [43]. On the other hand, GPT simply uses next word prediction and significant prior work demonstrates flexible applications of this style of modelling [37]. When there is alignment between the strong-inductive-bias FMs and task, we note the drastically improved performance with orders of magnitude fewer parameters (e.g., on 20News with bi-encoders). However, when there is misalignment, these FMs fail — inspecting T0-3B FM mispredictions on 20News, for 24.4% of examples, the model generated one of the valid classes for a particularly similar topic classification task called AGNews, which, unlike 20News, was part of the T0 training data. The issue is far more severe for T0-11B, which entirely fails on the task. This never happens for GPT mispredictions.

A challenge for enabling FOCUS is supporting finer-grained classification tasks. With GPT, prompting succeeds on the 2-way Sent140 classification and struggles on the 20-way 20News task. We observe a similar challenge using the CLIP models for 62-way F-EMNIST and construct a range of classification task granularities to study the trend. We create 4-way classification synthetics using subsets of 20News — Coarse requires classifying an input ∈ {politics, automobiles, baseball, or medicine}, while Fine requires classifying an input ∈ {politics, politics of guns, politics in the Middle East, or religion}. GPT-175B incurs a 55% drop from Fine to Coarse (Figure 3 (left)), and 40% performance drop on the longest vs. shortest textual examples (≤ 500 vs. ≥ 2500 characters) in Coarse. Further, zero-shot performance is quite uneven across classes on F-EMNIST and 20News (Figure 3 (right)) and across users with non-IID private distributions (Figure 7), suggesting FMs are not robust zero-shot.

4.3 To what degree does FOCUS allow personalizing to the user’s data distribution?

Personalization is a well-studied question in the FL literature [3, 20] and we similarly evaluate FOCUS along this axis. An ideal privacy framework should support users with zero-to-many private examples, and flexibly incorporate the
Figure 3: FMs struggle on fine-grained classification (left) and can provide highly uneven zero-shot performance across classes (right). Additional violin plots by user level accuracy are in Appendix 8.

Figure 4: (Left) Sent140 results including the “Public Prompt” baseline. (Right) Reddit and Sent140 results comparing 0-shot vs. k-shot performance by k, and comparing k-shot with examples selected from the user’s small labeled training dataset (“User Privacy”) vs. randomly from the aggregate training dataset of all users (“No User Privacy”). The accuracy is averaged across users.

different degrees of available user context. While FMs display strong out-of-the-box transfer abilities on certain tasks, each user often holds a few training examples per task in our setting — we consider users with \(<k\), \(\geq k\), and 0 train examples, for some small \(k\).

Users with few \(<k\) private examples. Building from the zero-shot prompting with task descriptions, we now turn to prompting with task demonstrations. We compare: randomly chosen training examples from the aggregate pool of all users’ training data (“No User Privacy”) and examples chosen from the few training examples the user has available (“User Privacy”). In contrast to Table 1, this experiment includes no task description in the prompt to focus on the effect of task demonstrations. We report results for \(K \in \{0, 3, 5\}\) on the Reddit and Sent140 benchmarks in Figure 4. In summary, (1) \(K > 0\) user-level task demonstrations provides improvements over the zero-shot baseline across model sizes, barring GPT-6.7B on Reddit. (2) Few-shot learning enables small FMs to match the performance of FMs with orders of magnitude more parameters — for example, on Sent140, comparing GPT-125M k-shot performance to GPT-6.7B 0-shot, and GPT-6.7B k-shot to GPT-175B 0-shot. (3) We find personal user level prompts provide larger gains across model sizes compared to the non-user level.

Users with enough \(\geq k\) private examples. Inference-only strategies tend to require large models and well-designed prompts — while effective for users with very few demonstrations, a framework with clean incentives should benefit users with relatively more data. In our setting, a subset of users have on the order of a few dozen to few hundred examples. As a proof of concept, using \(K = 16\) randomly chosen examples per class and tuning adapter networks on the order of 100k-1M parameters [29], we observe improvements over zero-shot performance to 41.3% \(\pm\) 6.2% on F-EMNIST (19% gain) and 71.3% \(\pm\) 0.3% on 20News (8% gain), which are both challenging zero-shot. Notably for 20News this quality is achieved using an order 100M parameter low-inductive bias language model instead of the bias bi-encoder model. Experimental details are in Appendix 7.

Users with \(k = 0\) private examples. Several personal tasks (e.g., spam classification, message generation) are of interest to many users. We suggest that a public prompt repository, or library of public k-shot examples, can be shipped down to users along with the FM. Accordingly, we evaluate “Public” third baseline beyond “No User Privacy” and

---

6Excitingly, relevant resources are already in development, originally proposed for alternate purposes [5].
Table 2: Cost comparison. Assumes 100 communication rounds, 32 batch size, 1000 steps, & 512 sequence length for FL; 1024 max sequence length for the 10B model. Assumes full precision.

| Resource     | FOCUS (10B model) | FEDERATED LEARNING (100M model) | Modern Phone [23] |
|--------------|-------------------|---------------------------------|-------------------|
| Communication| 40 GB, 1.5 hours download | 40 GB, 13 hours upload & download, barring latency | 61 Mbps download & 8 Mbps upload [2] |
| Training     | None              | 1e16 FLOPs                      | 16 TFLOPs, 4 GB RAM |
| Inference    | 2e13 FLOPs        | 1e11 FLOPs                      | 16 TFLOPs, 4 GB RAM |
| Storage      | 40 GB on disk     | 400 MB on disk                  | 1 TB disk         |

Analysis: FOCUS flexibly personalizes depending on the amount of user data. There are clear improvements from the “Public” to “No User Privacy” to “User Privacy” baseline, which is intuitive from a distribution shift perspective. Though it is exciting that FMs are sensitive to even these small peculiarities across users — it suggests that while these FMs may learn the core reasoning patterns during pretraining, they can actually incorporate personal information as well. We include an extended analysis of why user-level examples may help in Appendix 8. However, as shown in violin plots in Figure 7 FMs are not always robust zero-shot across all classes and users. The other exciting property of this paradigm is that just a few task demonstrations (either privately-owned or hosted in a public-store) enable using orders-of-magnitude more parameter-efficient models.

4.4 To what degree does FOCUS enable personalization to the user’s task distribution?

FOCUS is competitive in settings classically tackled with FL. Next, we suggest FOCUS enables personal tasks that are altogether incompatible with FL. For instance, Arora et al. [4], recently proposes the novel problem setting of open-domain multi-hop reasoning over data spanning multiple privacy scopes and identifies the need for an alternate privacy model beyond that provided by private FL. Further, we argue that unlike FL, FOCUS supports (1) idiosyncratic label schema — consider client i who performs our coarse synthetic vs. client j who performs fine, and (2) idiosyncratic reasoning patterns. For the latter, the reasoning patterns required by our evaluation benchmarks are relatively standard (e.g., “perform classification”): thus they are likely well-represented in the FM pretraining data and also easier to incentivize several users to participate in for FL. Consider a user who requires the idiosyncratic task: e.g., “Did I spend more time watching videos or working last month?” It is not obvious whether users want to expose participation in such tasks nor how many users seek such tasks. Figure 4 gives a toy demo of idiosyncratic tasks over a private email.

5 Systems Feasibility

Next, we examine the feasibility of each paradigm on modern hardware, focusing conservatively on personal phones. Though, we note that users have access to GPUs through a wide range of devices such as phones, personal computers, autonomous vehicles, and at-home servers. Enterprise users are even more well-resourced. For exposition, we take an order 100M parameter model for FL and 10B parameter model for FOCUS, based on the trends in Section 4.1.

The costs are summarized in Table 2 along the following dimensions: Communication FOCUS requires locally downloading and FL requires communicating the model repeatedly between users and the central server. Training requires \((2 \times 3 \times \text{model parameters} \times \text{steps} \times \text{batch size} \times \text{input length})\) FLOPs, and communication between clients and the central server, a function of network download/upload speeds and model size. FMs in FOCUS are frozen. Inference takes \((2 \times \text{model parameters} \times \text{input length})\) FLOPs, assuming key and value vectors for the attention computations are cached.

The RAM on modern phones is a key challenge along the above axes. Future-looking, we are optimistic, noting: (1) readily available support for offloaded training and inference [61, 59], (2) clear trends towards more RAM and effective small FMs [1, 69], and (3) some users own other capable devices beyond phones. As we freeze and perform inference with FMs, we can use quantization (i.e., 8 or even 4-bit precision) and pruning, to improve efficiency [26]. It is difficult to apply such techniques during training [10]. An exciting direction is, instead of requiring users to host FMs, developing methods for users to perform verifiably private inference with remote models [70].
**Task multiplicative factor** Critically, FL incurs these costs per personal task the client requires. A promise of FMs is their adaptability, enabling the use a single FM (order 100M-100B parameters) for multiple tasks versus a unique model per personal task in FL. Our work shows there is work to be done to achieve this promise, though we see the development of more versatile FMs over time [77, 69, 37]. For small FL models (100K parameters), the multiplicative factor is insignificant, yet model quality improves using more parameters, or training for longer over more data [32]. Accordingly, recent work in federated learning increasingly considers models on the order 20M-200M parameters [57, 62, 42, 28] and the economics are competitive in this regime.

**Incentives** Ultimately, users will have their own utility functions that take their privacy sensitivity, quality requirements, and systems constraints as inputs, determining the solution they adopt [19]. For example, for users who want or need perfect or close-to-perfect secrecy (i.e., in government or medical settings where data cannot be shared publicly), even a weak FM may be better than a high-performing FL method. Further, if the FM baseline is competitive even for a subset of the $N$ clients, the subset is disincentivized from participating. Not only do these clients avoid privacy leakage and communication costs of FL, but the fact that the zero-shot abilities succeed for these and not the other clients indicates there may be a distribution mismatch between the subsets of clients that make different decisions.

## 6 Conclusion

Amidst a recent focus on statistical notions of privacy, FOCUS suggests that perfect secrecy might be possible for a range of personal tasks. This is only a proof-of-concept and there are several future challenges such as the fragility of prompting, out-of-domain degradation, and slow runtime of performing inference using large models. Critically, there are numerous risks to using FMs. FMs hallucinate knowledge when uncertain, are predominantly available for resource-rich languages, and are expensive to pretrain. We release our implementation to help facilitate further study.

## Acknowledgements

We thank Sabri Eyuboglu, Neel Guha, Michael Zhang, Laurel Orr, Kawin Ethisayarajh, Deepak Narayanan, Mayee Chen, Maya Varma, Gary Cheng, Rohith Kuditipudi, Xuechen Li, Sidd Karamcheti, and Rishi Bommansani for their helpful feedback and discussions. We gratefully acknowledge the support of DARPA under Nos. FA86501827865 (SDH) and FA86501827882 (ASED); NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); the Moore Foundation, NXP, Xilinx, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Swiss Re, Brown Institute for Media Innovation, Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program, Fannie and John Hertz Foundation, National Science Foundation Graduate Research Fellowship Program, Texas Instruments Stanford Graduate Fellowship in Science and Engineering, and members of the Stanford DAWN project: Teradata, Facebook, Google, Ant Financial, NEC, VMware, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

## References

[1] 2022. [Blake’s ios device specifications grid](#)

[2] 2022. [United states’ mobile and fixed broadband internet speeds](#)

[3] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. In *arXiv:1912.00818v1*.

[4] Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2022. Reasoning over public and private data in retrieval-based systems.

[5] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Feyvy, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xingru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.
[6] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (PMLR).

[7] D. E. Bell and L. J. LaPadula. 1976. Secure computer system: Unified exposition and multics interpretation. The MITRE Corporation.

[8] Rishi Bommasani, Drew A. Hudson, E. Adeli, Russ Altman, Simran Arora, S. von Arx, Michael S. Bernstein, Jeanette Bohg, A. Bosselut, Emma Brunskill, and et al. 2021. On the opportunities and risks of foundation models. arXiv:2108.07258.

[9] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, , and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS).

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

[11] Sebastian Caldas, Sai Meher, Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, , and Ameet Talwalkar. 2019. Leaf: A benchmark for federated settings. In Workshop on Federated Learning for Data Privacy and Confidentiality.

[12] Ran Canetti. 2001. Universally composable security: A new paradigm for cryptographic protocols.

[13] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, , and Xueqi Cheng. 2021. Fedmatch: Federated learning over heterogeneous question answering data. In Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM ’21).

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

[15] Gregory Cohen, Saeed Afshar, Jonathan Tapson, , and André van Schaik. 2017. Emnist: an extension of mnist to handwritten letters.

[16] LEi Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. In PROCEEDINGS OF THE IEEE.

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

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference.

[19] Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

[21] Haokun Fang and Quan Qian. 2021. Privacy preserving machine learning with homomorphic encryption and federated learning. Future Internet.

[22] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension.

[23] Andrew E. Freedman. 2021. Apple a15 bionic powers iphone 13 and ipad mini.

[24] Alec Go, Richa Bhayani, , and Lei Huang. 2009. Twitter sentiment classification using distant supervision.

[25] Andy Greenberg. 2017. How one of apple’s key privacy safeguards falls short.

[26] Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In ICLR 2016.

[27] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annanvaram, and Salman Avetisianh. 2020. Fedml: A research library and benchmark for federated machine learning. arXiv preprint arXiv:2007.13518.
[28] Agrin Hilmkil, Sebastian Callh, Matteo Barbieri, Leon Rene Sutfeld, Edvin Listo Zec, and Olof Mogren. 2021. Scaling federated learning for fine-tuning of large language models. In arXiv:2102.00875.

[29] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In Proceedings of the 36th International Conference on Machine Learning (PMLR).

[30] Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

[31] Vincent C. Hu, David F. Ferraiolo, and D. Rick Kuhn. 2006. Assessment of access control systems. NIST.

[32] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. https://arxiv.org/abs/2001.08361.

[33] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning.

[34] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In arXiv cs.CL 2107.13586.

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, , and Xiaou Tang. 2015. Deep learning face attributes in the wild.

[36] Rabeek Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathiasi, Veselin Stoya, and Majid Yazdan. 2022. Perfect: Prompt-free and efficient few-shot learning with language models. In Annual Meeting of the Association for Computational Linguistics.

[37] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).

[38] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. ICLR.

[39] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In Proceedings of 40th IEEE Symposium on Security & Privacy (S&P).
[75] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In ICML.

[76] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In Advances in Neural Information Processing Systems 32.

[77] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Ayeek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. arXiv.

[78] Yufeng Zhan, Jie Zhang, Zicong Hong, Leijie Wu, Peng Li, and Song Guo. 2021. A survey of incentive mechanism design for federated learning. IEEE Transactions on Emerging Topics in Computing.

[79] Tony Z. Zhao, 1 Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In arXiv:2102.09690v2.

[80] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-iid data. arXiv:1806.00582.
Privacy Benchmark | Train Examples per User | Total Test Set Size
--- | --- | ---
Reddit (Non-IID) | 34.1 | 25.5k
MRQA (Non-IID) | 501.5 | 3.0k
Sent140 (Non-IID) | 2.4 | 286.6k
20News (IID) | 113.1 | 7.5k
CelebA (IID) | 21.4 | 11.1k
CIFAR-10 (IID) | Varied | 10.0k
F-EMNIST (IID) | 226.8 | 81.7k

Table 3: We show statistics for each benchmark and whether each has IID or Non-IID data per user.

7 Appendix: Experimental Details

7.1 Benchmark Protocols

We release our code for reproducability and future work, including task and class descriptions we use, prompt formats, and task scoring functions. The benchmarks were originally proposed in Caldas et al. [11], Lin et al. [42] and are openly-accessible. Benchmark size-statistics are shown in Table 3 and here we provide details of any dataset sampling or prompting choice.

CelebA This is an image classification benchmark and we evaluate using the CLIP-ViT32B model using the full test dataset. We encode the binary class descriptions and match the image embeddings to the closest class embedding.

task_description = NA
class_descriptions = ["frowning", "smiling"]

CIFAR-10 This is an image classification benchmark and we evaluate using the CLIP-ViT32B model using the full test dataset. We encode the class descriptions and match the image embeddings to the closest class embedding.

task_description = NA
class_descriptions = ["airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", "truck"]

F-EMNIST This is an image classification benchmark and we evaluate using the CLIP-ViT32B model using the full test dataset. We encode the class descriptions and match the image embeddings to the closest class embedding. Note that the CLIP models are not case-sensitive.

task_description = NA
class_descriptions = ["The picture is of the uppercase letter <<c>>", "The picture is of the lowercase letter <<c>>", "The picture is of the digit <<c>>"] where the classname c gets inserted as <<c>>

Sent140 This is a text classification benchmark and we evaluate using the GPT, T0, and bi-encoder model variants, using a constant random subsample of 2.5% of clients (7.2k examples), due to cost restrictions. We include clients with $\geq 0$ labeled training examples. For the autoregressive models, we map the output to a class using the first generated token (scoring code is released). For the bi-encoder model, we encode class descriptions “positive” and “negative” and match the input embeddings to the closest class embedding. The fixed public prompt used in Figure 4 is shown below:

task_description = "Is this text positive or negative? Text: "
class_descriptions = ["positive", "negative"]

20News This is a text classification benchmark and we evaluate using the GPT, T0, and bi-encoder model variants, using the full test dataset. For the autoregressive models, we map the output to a class using the first generated token (scoring code is released). For the bi-encoder model, we directly encode class descriptions as follows and match the input embeddings to the closest class embedding.

https://github.com/FedML-AI/FedNLP
https://github.com/TalwalkarLab/leaf
**Synthetic 20News** This is a text classification synthetic dataset constructed by subsampling 1k points from the 20News dataset that have one of the desired labels. We evaluate using GPT-6.7B and GPT-175B. We specifically create two synthetics, termed “Coarse” and “Fine”. The Coarse details are as follows:

```python
task_description = "Is the topic politics, baseball, medical, or automobiles?"
class_descriptions = ["politics", "baseball", "medical", "automobiles"]
```

The Fine details are as follows:

```python
task_description = "Is the topic politics, religion, politics guns, or politics middle east?"
class_descriptions = ["politics", "religion", "politics guns", "politics middle east"]
```

**MRQA** This is a reading comprehension benchmark and we evaluate using GPT model, using the full test dataset. Given an input question and context, we prompt the model as follows:

```python
task_description = "What is the answer span to the following question?"
```

```python
Question: When was Barack Obama born?  
Answer: 1983
```

```python
class_descriptions = NA
```

**Reddit** This is a language modeling benchmark and we evaluate using the Grover and GPT variants, using the full test dataset. However, for GPT-6.7B and GPT-175B we subsample 5% of the clients.

```python
task_description = NA
class_descriptions = NA
```

### 7.2 Experimental Details

**Inference-only strategies** For the results in Table 1 and Figure 2 we use the models zero-shot. For autoregressive models, we simply prompt the model with the exact task descriptions specified above. For similarity search, we use the exact class descriptions specified above. For all results in Figure 4 we use no task description to focus on the effect of task demonstrations. Note that to hold the number of task demonstrations per user constant across baselines, the “No User Privacy” baseline randomly selects the minimum of \{number of user’s training examples, \( k \)\} for \( k \in \{3, 5\} \) — for example on Sent140, the average number of task demonstrations per user is 2.4, and the number of in-context examples available for the “User Privacy” baseline is the user’s training data size.

**Lightweight fine-tuning** For these experiments, we randomly selected 16 examples per class, from the aggregate training data of all users, for training.

For the vision tasks, we use the CLIP-ViT32B base FM and the adapter network is a 2-layer MLP with ReLU and batch normalization. This results in 130K trainable parameters. We use the following hyperparameters and report results over
three random seeds: 64 hidden dimension, 50 max epochs, 32 batch size, 1e-3 learning rate, 0.5 weight decay, and 0.9 momentum. These parameters performed best of hidden dimensions $\in \{64, 128\}$, weight decay $\in \{0.5, 0.9\}$, and learning rate $\in \{1e-3, 1e-4\}$.

For the language tasks, we use PERFECT, a recent adaptation method from Mahabadi et al. [48] which inserts an adapter layer after the feed-forward block of each transformer layer of a RoBERTa-Large model, an early FM consisting of 355M parameters. This results in 3.3M trainable parameters. We train using the default hyperparameters and report results over three random seeds.

7.3 Additional Discussion of Baselines

Given the heterogeneity of FL evaluations [27], we aim to compare vanilla baselines for both paradigms in this first investigation. For baselines within the FOCUS and FL frameworks, we consider advanced neural architectures with standard training and inference methods. For FOCUS, this involves off-the-shelf, publicly accessible models with manual prompting, without any tuning. For FL, we include numbers that use the most advanced architecture reported for the benchmark in prior work (to the best of our knowledge), using standard FedAvg for training and none of the following factors which significantly degrade quality.

- **Differential privacy** Differential privacy (DP) seeks to provide a guarantee that we cannot reconstruct or memorize the sensitive examples appearing in our training data. The basic approach to achieve this involves adding noise to the data during preprocessing, gradients during training, or elsewhere in the ML pipeline, where the amount of noise depends on the desired privacy parameters. DP faces a tradeoff between stronger privacy protection (achieved by adding more noise) and better convergence performance [71]. Our FL baselines applied no DP.

- **Adversarial users** Adversaries can execute a data poisoning attack (i.e. poison the labels of the contributed data), send random updates to the central server, and/or execute a model replacement attack (i.e. manipulate the shared model to enforce that it performs a desired subtask, while maintaining performance on the original FL task) [6, 74]. As the fraction of adversaries increases, the quality of the FL model tends to degrade [39].

- **Adversarial central server** Because the central server is a single point of failure and attacks upon the central server significantly compromise privacy, decentralized federated learning has emerged as an alternative line of work. However, this again degrades quality [68].

- **Communication efficiency** Popular methods to reduce the communication cost of FL such as compressing the gradients passed between client and server tend to degrade quality [62].

- **Non-IID Data** Note that some of our benchmarks are already non-IID. Under SGD, we require an unbiased estimate of the full gradient. This relies on computing the stochastic gradients over IID samples. User data is likely non-IID, which can degrade the quality and increase the convergence time of the FL model trained with FedAvg [80]. However, recent architectures appear more robust to heterogeneous data under FL [57].

Similarly, for the FM baselines, we try the simplest possible adaptation techniques in the space and report numbers across a range of models and model sizes. Future work could strengthen the FOCUS baselines using advanced prompting strategies. For example, we use unconstrained answer spaces, whereas classification tasks are compatible with constrained spaces (i.e. rather than having the autoregressive model generate any token in the vocabulary, restrict to tokens that correspond to valid classes such as “positive” and “negative” for sentiment classification). Other example strategies include using continuous prompts, learned task templates, and calibration to reduce model bias towards certain answers [46, 79, 37].

7.4 Examples of each few-sample adaptation technique

Using a topic classification task as an example, the following demonstrate examples of (1) prompting with task descriptions or labeled examples, (2) zero-shot classification via nearest neighbors similarity search, and (3) FM tuning.

An example of prompting with zero-to-few task demonstrations follows. The above prompt includes the task description and the below prompt includes task demonstrations.

```java
// With Task Instructions (Zero Shot)
prompt = "Is the text about medicine, baseball, atheism, electronics, ..., or hockey? Text: I believe the Cubs have the best record ever in the MLB.
```

[https://github.com/facebookresearch/perfect](https://github.com/facebookresearch/perfect)
prediction = model.generate(prompt)
// With Task Demonstrations (Few Shot)
prompt = "Is the text about medicine, baseball, atheism, electronics, ..., or hockey?

Input: They detect the oscillator operating in the detector.
Label: electronics

Input: I believe the Cubs have the best record ever in the MLB.
Label: 

prediction = model.generate(prompt)

>> print(prediction)
>> "baseball"

An example of similarity search for zero-shot classification follows.

e.example = "I believe the Cubs have the best record ever in the MLB."

class_descriptions = ["baseball", "hockey", "politics in the middle east", "electronics", "medicine", "automobiles", "atheism", "cryptography"]

embedded_input = model.encode_query(example)
embedded_descriptions = model.encode_keys(class_descriptions)

// predict label with highest embedding similarity score to the example
prediction = max_similarity_class(encoded_input, encoded_descriptions)

>> print(prediction)
>> "baseball"

The setup for lightweight tuning follows. Note that the adapter approach may be replaced by other lightweight tuning options such as in Li and Liang [38], Hu et al. [39].

// train a model on (x, y) pairs with stochastic gradient descent
training_data = [
  ("Does the Proventil inhaler fall into the steroid category?", "medicine"),
  ("It is called phimosis, usually it is due to an inflammation", "medicine"),
  ("Please give me the address for Texas Ranger ticket orders.", "baseball"),
  ("I need the 84 boxscores of an NHL team for personal research", "hockey"),
  ("Giants Win the Pennant!! Giants Win the Pennant !!", "baseball"),
  ("Where can I get a New York taxi?", "automobile"),
  ...
]

model.freeze_parameters() // freeze non-adapter parameters
model.train(training_data)

example = "I believe the Cubs have the best record ever in the MLB."
prediction = model.predict(example)

>> print(prediction)
>> "baseball"
8 Appendix: Additional Results and Analysis

User vs. Non-User Level Task Demonstrations  In Section 4.1, we observe that the “User Privacy” prompts provide larger performance gains compared to the “No User Privacy” prompts. We hypothesize that (1) label distributions user level task demonstrations help because user labels are generally from a subset of classes of the overall task, and (2) input distributions user level task demonstrations help due to linguistic or topical similarity between user demonstrations.

Towards the former, on a per-user basis, as the entropy of the training data labels decreases, the accuracy tends to increase, as in Figure 5. The entropy is computed for users with at least 4 training examples and each point represents at least 200 examples. As the entropy of the user-level train set increases, the improvement of the “User Privacy” baseline over the “No User Privacy” baseline tends to decrease, suggesting this is an important reason for the difference, though there is still some gap between the baselines for high entropy points. Towards the latter, qualitatively certain users clearly write with specific artifacts or about certain topics, though overall the examples are quite noisy. Three users for whom this occurs are in Table 4.

Idiosyncratic tasks  Figure 6 is a toy demo that involves taking a private email from the dataset in Arora et al. [4] and performing idiosyncratic tasks over the email. Aligning incentives and executing FL for every personal task may be challenging, and FOCUS may provide an alternate path forward. However, as shown in the Figure and benchmark experiments, zero-shot performance does not always suffice out-of-the-box using simple task descriptions.

Violin Plots  Figures 7 and 4 provide violin plots with respect to zero-shot performance across tasks. These plots demonstrate that on tasks where user data are non-IID, the model performs unevenly in zero-shot use across different groups. Further, on the fine-grained classification tasks, performance is uneven across classes. For example, on F-EMNIST, the model often either always selects the uppercase or lowercase version of a letter, resulting in close to 100% accuracy on the selected case and 0% on the unselected case. On 20News, accuracy by class is above 50% for every class except for “politics”, where the accuracy is 12.9% and “religion”, where the accuracy is 9.6%. We suggest this is because “politics guns” and “politics in the middle east” related to politics, and “Christianity” and “atheism” related to religion are other classes in the 20News task — using the inference-only strategies based on natural language descriptions of these classes, it is difficult to express the distinction between these similar classes. We require additional
| User | Task Demonstrations |
|------|---------------------|
| User 1 (Sent140, Syntax) | *Example 1* @MYIDOLTOWN Oh my gosh!!!!!!!!!!!!!!! I’m speechless where did you see/hear about this story?  
*Example 2* Just updated my MySpace profile am about to send Sweet Danny a comment on his MySpace page going to sleep shortly zzzzzzz!!!!!

*Example 3* Had fun happy day going to sleep for Church; Sunday School in the morning

*Example 4* Am doing washând as always!listening to Danny!!!!!!!!!!!!!!!!!

*Example 5* @dannygokey Happy to hear you’re working on Sophia’s Heart I’ll always support you; Sophia’s Heart!!!!!!!!!!!!!! |

**Test Input** @Cupcake1012 I <3 the Gokey Gang; Danny always; forever Danny has the best fans ever!!!!!!!!!!!!!!!!!!!!!!!!!!!! |

| User 2 (Sent140, Artifacts) | *Example 1* ... A kidney stone. Really?? Ugh [http://tinyurl.com/qsw9vq](http://tinyurl.com/qsw9vq)  
*Example 2* Ahhh everything hurts... warming up my bed buddy and going back to sleep. [http://tinyurl.com/ol4ugp](http://tinyurl.com/ol4ugp)  

*Example 3* Feelin so sick its dumb. [http://tinyurl.com/o6glq2](http://tinyurl.com/o6glq2)  
*Example 4* I got so much done today!! Got my car fixed, went to the bank. Now im at work to make monies [http://tinyurl.com/of2ane](http://tinyurl.com/of2ane)  

*Example 5* name the movie... “i wanna be like yooo-oo-ooo...”; [http://tinyurl.com/pqbyst](http://tinyurl.com/pqbyst) |

**Test Input** i wish i was a snail [http://tinyurl.com/pgysd5](http://tinyurl.com/pgysd5) |

| User 3 (Reddit, “Justice” topic) | *Example 1* dion lewis stiff arm plesse my favorite professor in law school actually argued at scotus  
*Example 2* scotus for the petitioner in mccleskey i hope for his  

*Example 3* his sake (and for the sake of justice) that  
*Example 4* that they revisit it at some point |

**Test Input** ... am a personal injury attorney <PAD>

Table 4: Examples reflecting consistent linguistic patterns, artifacts, and topics across a user’s examples.

investigation into the zero-shot robustness of FMs, and effective ways of describing classes that avoid class embeddings that point in the same or similar directions in embedding space.
Idiosyncratic Personal Tasks

Email: --Original Message -- From: Steffes, James D. 
Sent: Monday, July 30, 2001 7:36 AM To: Demehl, 
Ginger Cc: Shapiro, Richard Subject: FW: EES-USA 
Ginger – Please forward to Rick's Government Affairs 
team in the USA. By the way, this is better if printed 
on legal paper. Thanks, Jim --Original Message -- 
From: Leff, Dan Sent: Friday, July 27, 2001 10:34 AM 
To: Lavorato, John; Belden, Tim; Presto, Kevin M.; 
Richter, Jeff; Aucoin, Berney C. Perry, Todd; Steffes, 
James D.; Beck, Sally; Colwell, Wes; Apollo, Beth; 
Haedicke, Mark E.; Bradford, William S.; Herndon, 
Rogers; Black, Don Cc: Dietrich, Janet Subject: EES- 
USA Attached is a diagram, which functionally lays out 
EES - USA as a follow up to our offsite of July 18. 
Please review and let Janet, Dave or me know if you 
would like to discuss Thank you again for your 
participation in the meeting. We are making progress! 
Thanks – Dan

Answer: 

User task description: What tasks are assigned to me in this email?
GPT3: The tasks assigned to the recipient of this email are to review the attached

User task description: What is the timeline in this email?
GPT3: The timeline in this email is from July 27, 2001 to July 30

User task description: Clean the email
GPT3: Hi Ginger, Please find the attached file from Jim Steffes regarding the EES-USA. He suggests that it would be better if printed on legal paper. Thanks, Dan

User task description: What is EES-USA?
GPT3: EES-USA is a subsidiary of EES International, a leading provider of energy efficiency and sustainability solutions. EES-USA provides a range of services including energy efficiency consulting, project management, and financing for commercial, industrial, and residential customers in the United States ...

Figure 6: A toy demo of a user executing idiosyncratic personal tasks over a private email, using a natural language interface and FM inference.
Figure 7: Violin plots by user id. We note that the Sent140, MRQA, CelebA, and Reddit contain Non-IID data splits across users and F-EMNIST and 20News use IID mappings of data to clients. Non-IID plots appear more multi-modal.