A Survey on Federated Fine-Tuning of Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) have demonstrated impressive success across various tasks. Integrating LLMs with Federated Learning (FL), a paradigm known as FedLLM, offers a promising avenue for collaborative model adaptation while preserving data privacy. This survey provides a systematic and comprehensive review of FedLLM. We begin by tracing the historical development of both LLMs and FL, summarizing relevant prior research to set the context. Subsequently, we delve into an in-depth analysis of the fundamental challenges inherent in deploying FedLLM. Addressing these challenges often requires efficient adaptation strategies; therefore, we conduct an extensive examination of existing Parameter-Efficient Fine-tuning (PEFT) methods and explore their applicability within the FL framework. To rigorously evaluate the performance of FedLLM, we undertake a thorough review of existing fine-tuning datasets and evaluation benchmarks. Furthermore, we discuss FedLLM's diverse real-world applications across multiple domains. Finally, we identify critical open challenges and outline promising research directions to foster future advancements in FedLLM. This survey aims to serve as a foundational resource for researchers and practitioners, offering valuable insights into the rapidly evolving landscape of federated fine-tuning for LLMs. It also establishes a roadmap for future innovations in privacy-preserving AI. We actively maintain a GitHub repo to track cutting-edge advancements in this field.

1 Introduction

Large Language Models (LLMs), such as GPT-4o (OpenAI, 2024), DeepSeek-R1 (Guo et al., 2025), and Qwen3 (Yang et al., 2025) have exhibited extraordinary proficiency across a spectrum of downstream tasks. These LLMs, distinguished by their ability to capture complex semantic knowledge, have established new performance benchmarks in computational linguistics. However, despite their impressive capabilities, LLMs cannot be directly deployed for specific downstream tasks without appropriate adaptation (Hu et al., 2021). Furthermore, training LLMs directly on downstream task datasets presents substantial challenges. The massive scale of model parameters leads to significant computational overhead (Tian et al., 2024b), while the scarcity of task-specific data constrains effective model training and increases the risk of overfitting. For example, training LLaMA2-65B involves processing approximately 1.4 trillion tokens, requiring 21 days of computation on on 2,048 NVIDIA A100 GPUs (Touvron et al., 2023a). Consequently, fine-tuning pre-trained LLMs has become the dominant paradigm (Dodge et al., 2020), enabling more efficient adaptation of LLMs to specific tasks while preserving their foundational knowledge acquired during pre-training.

The current mainstream LLM fine-tuning paradigms can be categorized into three approaches: 1) **Centralized Fine-Tuning** (as shown in Figure 1(a)): This approach aggregates local datasets from all data owners (clients) and uploads them to a central server for fine-tuning (Zhang et al., 2023e; Ding et al., 2023b). Despite its effectiveness, this approach raises significant privacy concerns (Wang et al., 2023b; Ye et al., 2024a; Tam et al., 2024) and is often impractical in real-world scenarios due to legal restrictions (e.g., GDPR (Voigt & Von dem Bussche, 2017)), which limit the centralization of sensitive personal data. 2) **Local Fine-Tuning** (as shown in Figure 1(b)): In this paradigm, each data owner fine-tunes the LLM locally using their private dataset. While this approach preserves data privacy, the limited size and diversity of local datasets often result in suboptimal model performance. For instance, models refined through local fine-tuning demonstrate

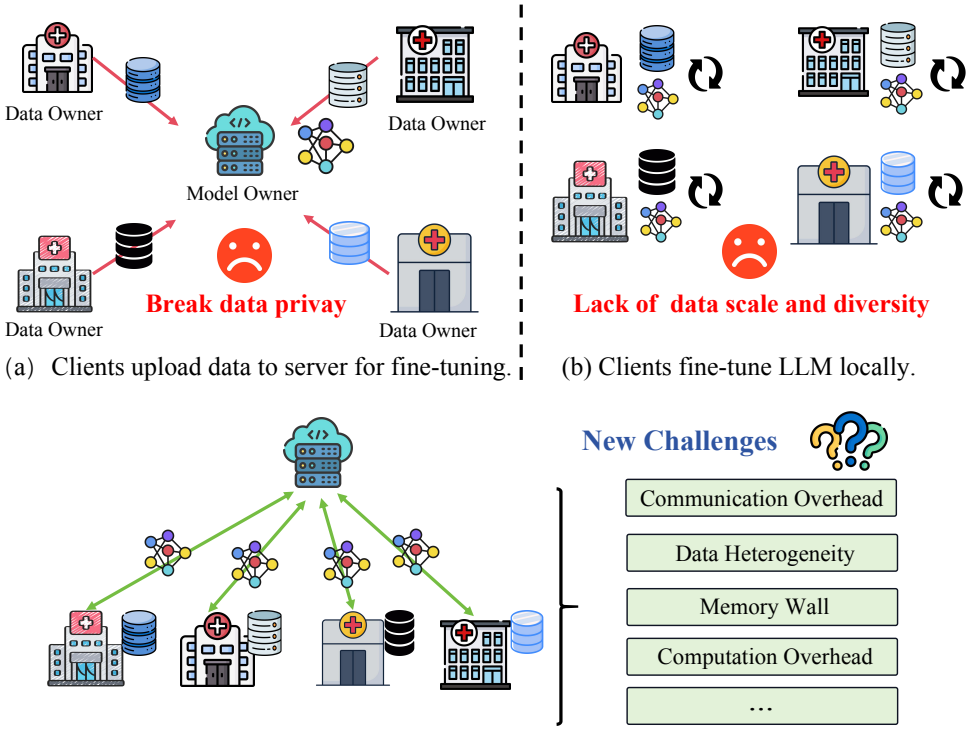

Figure 1: Illustration of three LLM fine-tuning paradigms: (a) **Centralized Fine-tuning**, where data is aggregated at a central server; (b) **Local Fine-tuning**, where models are trained independently on private datasets; and (c) **Federated Fine-tuning**, where model updates are shared while keeping data local.

a substantial performance degradation of up to 7% on the MMLU benchmark (Hendrycks et al., 2020) when compared to federated fine-tuning (Ye et al., 2024b). 3) **Federated Fine-Tuning** (as shown in Figure 1(c)): This approach enables collaborative model improvement while preserving data privacy by allowing clients to train the model locally and only sharing model updates with the central server (Li et al., 2025; 2023c). The server aggregates these updates to construct a global model, which is subsequently redistributed to clients for further refinement. This method addresses both the privacy concerns of centralized fine-tuning and the limited data diversity issues of local fine-tuning, making it a promising paradigm for adapting LLMs to specific downstream tasks (Xu et al., 2023e).

Despite these benefits, federated fine-tuning encounters several unique challenges, which significantly limit the effective deployment of FedLLM in real-world scenarios: 1) **Communication Overhead**: LLMs typically contain billions of parameters, such as LLaMA2-7B. Therefore, uploading these massive model parameters in each training round incurs substantial communication overhead, resulting in severe communication latency and excessive bandwidth requirements (Li et al., 2022a). 2) **Data Heterogeneity**: Data across participating clients exhibits substantial variation in both quality and statistical distribution (Ning et al., 2024). This Non-IID (Non-Independent and Identically Distributed) nature of federated data can introduce significant biases into model updates, leading to weight divergence, slower convergence rates, and ultimately compromised model performance (Fu et al., 2024a; Tian et al., 2024d). 3) **Memory Wall**: Participating clients, especially edge devices, generally possess limited available memory resources (Wu et al., 2024h;i; Zhan et al., 2024; Li et al., 2023g), which insufficiently support memory-intensive LLM fine-tuning. This memory wall fundamentally limits clients' effective participation in the federated fine-tuning process, preventing them from contributing valuable data to the global model and ultimately compromising model performance. 4) **Computation Overhead**: The hardware processing capabilities of participating clients are often limited (Wang et al., 2019a), making it challenging to meet the high computational demands of fine-tuning LLMs. This computational bottleneck substantially increases local training time and consequently prolongs

Table 1: **Overview of related surveys.** This comparison highlights whether each work addresses data privacy, aligns with the scope of LLMs, emphasizes efficiency, proposes evaluation benchmarks, and discusses applications and future directions.

Prior Surveys	Privacy	LLM	Efficiency	Benchmark	Application	Future Direction
Xu et al. (2024c)	✗	✓	✓	✗	✓	✓
Zhao et al. (2023b)	✗	✓	✓	✓	✓	✓
Gao et al. (2024b)	✗	✓	✗	✓	✗	✓
Li et al. (2024d)	✗	✓	✗	✓	✓	✓
Zheng et al. (2023b)	✗	✓	✗	✓	✓	✓
Han et al. (2024)	✗	✓	✓	✗	✓	✓
Xin et al. (2024)	✗	✓	✓	✓	✓	✓
Huang et al. (2024b)	✓	✗	✗	✓	✗	✓
Ye et al. (2023)	✓	✗	✗	✗	✗	✓
Jiang et al. (2024b)	✓	✗	✗	✗	✓	✗
Yuan et al. (2024a)	✓	✗	✗	✗	✓	✓
Chen et al. (2024c)	✓	✗	✗	✓	✓	✓
Chai et al. (2024)	✓	✗	✗	✓	✗	✓
Feng et al. (2023b)	✓	✗	✗	✓	✓	✓
Zhang et al. (2024j)	✓	✗	✗	✗	✗	✓
Yao et al. (2024)	✓	✓	✓	✗	✓	✓
Li et al. (2024e)	✓	✓	✓	✗	✓	✓
Zhuang et al. (2023)	✓	✓	✗	✗	✓	✓
Woisetschläger et al. (2024)	✓	✓	✓	✗	✗	✓
Ren et al. (2024)	✓	✓	✓	✗	✓	✓
Ours	✓	✓	✓	✓	✓	✓

the overall process. The extended training cycles reduce system efficiency and significantly increase energy consumption on resource-constrained devices, potentially deterring client participation.

To address these challenges, researchers have applied various Parameter-Efficient Fine-Tuning (PEFT) methods to FL, which can be broadly classified into five main categories: LoRA-based tuning (Hu et al., 2021; Tian et al., 2024c), prompt-based tuning (Lester et al., 2021), adapter-based tuning (Houlsby et al., 2019), selective-based tuning (Zaken et al., 2021), and other tuning methods (Shin et al., 2023; Chen et al., 2017). The core idea behind these methods is to minimize the number of trainable parameters by focusing on small, task-specific adjustments rather than fine-tuning the entire model. By updating only a subset of parameters or modifying model inputs (e.g., prompts), these approaches significantly reduce the communication overhead, computational burden, and memory usage of model fine-tuning, all while maintaining the performance of LLMs across diverse tasks.

Prior Surveys. Despite the continuous development of innovative federated fine-tuning methods, a significant gap remains in the comprehensive evaluation and comparison of these techniques. While existing surveys on FL provide valuable insights, they are typically limited to either traditional small-model FL settings or fail to offer a detailed analysis and evaluation benchmark specifically for federated fine-tuning of LLMs. Concurrently, although several surveys on PEFT have been proposed, these works predominantly focus on centralized fine-tuning scenarios, overlooking the unique challenges that arise when adapting such techniques to distributed, privacy-preserving settings. A detailed comparison of existing surveys is summarized in Table 1. To fill this gap, our survey aims to be the *first* to present a systematic examination of federated fine-tuning for LLMs, providing a thorough understanding of their evolution, effectiveness, and practical implementation challenges, along with standardized evaluation benchmarks that enable fair comparison across different approaches. This thorough analysis serves as a foundation for researchers and practitioners seeking to navigate the rapidly evolving landscape of federated LLM fine-tuning.

Contribution. This paper presents a comprehensive survey on the federated fine-tuning of LLMs. In contrast to existing surveys, the main contributions of this work can be summarized as follows:

1. We provide an exhaustive review of all relevant papers on federated fine-tuning up to date, offering an extensive analysis of the state-of-the-art techniques and their evolution in this field.

2. We conduct a detailed analysis of the key challenges in federated fine-tuning and propose a systematic taxonomy based on different fine-tuning approaches, including LoRA-based, prompt-based,

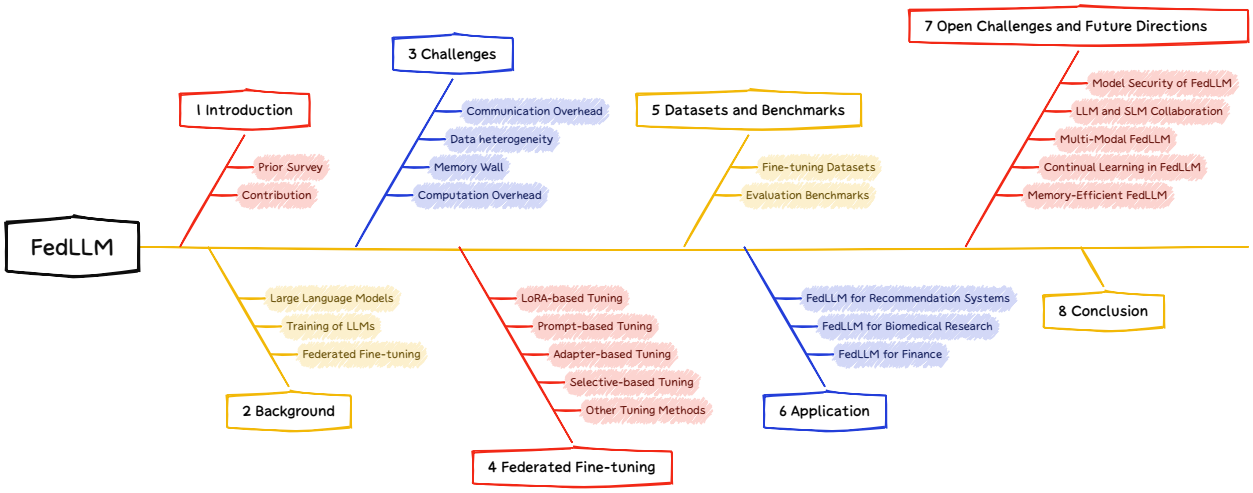

Figure 2: Overall structure of the survey.

adapter-based, selective-based, and other emerging tuning methods. We further provide an in-depth discussion of the advantages, limitations, and applicability of these methods.

3. We establish a comprehensive evaluation framework for federated fine-tuning of LLMs, encompassing fine-tuning datasets and evaluation benchmarks across diverse domains, while systematically analyzing and discussing diverse real-world application scenarios.

4. Finally, we outline promising research directions in FedLLM, aiming to guide future investigations toward more efficient, scalable, and privacy-preserving solutions that bridge the gap between theoretical advances and practical deployments in resource-constrained federated environments.

Figure 2 illustrates the organizational structure of this survey. Section 2 introduces the relevant background and fundamental concepts of LLMs and federated fine-tuning. Section 3 systematically examines the technical challenges and inherent limitations in the federated fine-tuning of LLMs. In Section 4, we present a comprehensive review of state-of-the-art federated fine-tuning techniques and methodologies. Section 5 presents representative fine-tuning datasets and evaluation benchmarks across various domains, specifically curated to assess the performance of federated fine-tuning in diverse scenarios. Section 6 explores practical applications of FedLLM. Section 7 outlines promising research directions, while Section 8 synthesizes key insights from this survey to inform and guide future research in this rapidly evolving field.

2 Background

2.1 Large Language Models

Large Language Models (LLMs) have demonstrated unprecedented capabilities across a wide range of natural language processing tasks, including machine translation (Wang et al., 2022a), text generation (Yu et al., 2022), sentiment analysis (Wankhade et al., 2022), and question answering (Zhu et al., 2021). Their exceptional performance stems from their remarkable ability to encode complex linguistic patterns, capture long-range contextual dependencies, and learn rich semantic representations. These capabilities have not only enabled LLMs to achieve state-of-the-art results on a broad spectrum of academic benchmarks, but have also fueled transformative advances in real-world applications such as conversational AI, legal document analysis, medical decision support, and automated content generation.

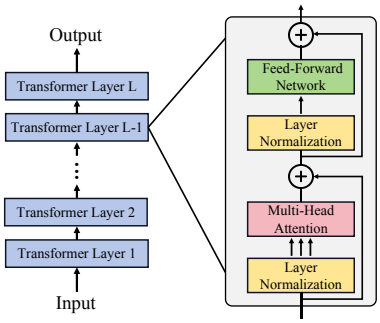

Figure 3: Architecture of LLMs.

Architecturally, modern LLMs are typically constructed by stacking dozens or even hundreds of transformer layers, where each layer incrementally refines the input through deep contextualization and abstraction. For example, LLaMA2-7B comprises 32 transformer layers stacked sequentially to capture hierarchical linguistic features. This deep, layered architecture enables the model to effectively integrate both local and global contextual information over long sequences, which is essential for complex language understanding tasks. Figure 3 illustrates the schematic structure of a prototypical LLM, where transformer layers are arranged in a vertically stacked fashion. Each transformer layer consists of two fundamental components: Multi-Head Attention (MHA) and Feed-Forward Network (FFN). Formally, the input to the l-th transformer layer is denoted as $h_{l-1} \in \mathbb{R}^{n \times d}$, where n is the sequence length and d is the hidden dimension of the model. The computational process within the l-th layer can be expressed as follows:

$$h_i^{'} = \text{MHA}(\text{LN}(h_{i-1})) + h_{i-1}, \tag{1}$$

$$h_i = \text{FFN}(\text{LN}(h_i^{'})) + h_i^{'} \tag{2}$$

where $\text{LN}(\cdot)$ represents layer normalization, which stabilizes the training dynamics by standardizing the activations, and $h_i^{'}$ denotes the intermediate activations after being processed by the MHA module.

2.2 Training of LLMs

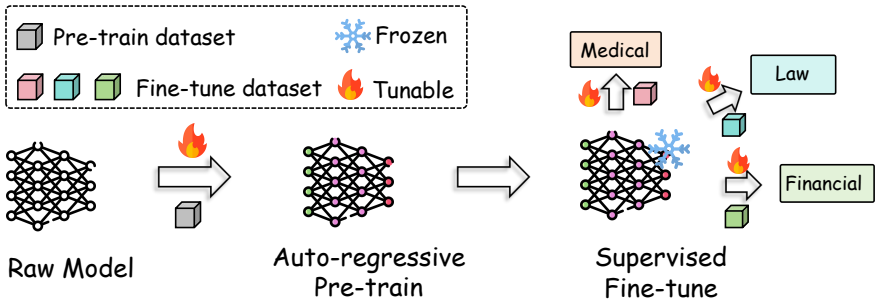

Figure 4: Schematic illustration of the two-stage LLM training process: 1) auto-regressive pre-training on large-scale corpora to develop general linguistic capabilities, followed by 2) supervised fine-tuning to align model outputs with specific task requirements or human preferences.

The training of LLMs encompasses two distinct stages (Xin et al., 2024): pre-training and fine-tuning, as illustrated in Figure 4. 1) **Pre-training** involves training the model on massive unlabeled text corpora (billions to trillions of tokens) drawn from diverse sources such as academic papers, websites, and books. This stage generally adopts the auto-regressive modeling (Yang et al., 2019) approach that predicts each token based on its previous context. Through extensive training, LLMs develop a wide range of capabilities, from basic semantic understanding to advanced reasoning across diverse domains. While computationally intensive, this unsupervised learning process builds robust and transferable representations that serve as a powerful foundation for various downstream tasks (Naveed et al., 2023).

2) The function of **fine-tuning** is to adapt the pre-trained model to specific downstream tasks through additional training on task-specific datasets (Ding et al., 2023b). This stage typically utilizes supervised learning to optimize the model's performance for particular applications. While fine-tuning is highly effective at specializing the model's general language understanding for targeted tasks, traditional fine-tuning methods often require centralizing data from various sources on a central server (Huang et al., 2025), which raises significant privacy and security concerns. These concerns have sparked growing interest in privacy-preserving fine-tuning paradigms, which seek to retain the benefits of model specialization while ensuring that sensitive user data remains decentralized and secure throughout the training process.

2.3 Federated Fine-tuning

Federated fine-tuning (Zhang et al., 2024b; Yi et al., 2025) has emerged as a promising paradigm for adapting LLMs to specific downstream tasks while preserving data privacy. Unlike conventional centralized approaches

that require sensitive data aggregation at a central server, federated fine-tuning enables distributed clients to adapt LLMs on local private datasets, sharing only model updates with the coordinating server. This privacy-preserving approach aligns well with modern data protection requirements and user expectations. However, the massive scale of LLM parameters and heterogeneous data distributions across clients introduce significant technical challenges. These challenges encompass prohibitive communication bandwidth requirements for transmitting model updates, convergence difficulties when training across heterogeneous data distributions, excessive memory demands that strain client-side resources, and intensive computational overhead that impacts efficiency and energy consumption. To address these challenges, researchers have proposed various parameter-efficient federated fine-tuning approaches, each strategically designed to mitigate resource constraints while maintaining model performance on downstream tasks. These innovative methods can be broadly categorized as follows:

- 1) LoRA-based Tuning (Hu et al., 2021): This methodology leverages the intrinsic low-rank nature of weight updates by decomposing them into low-rank approximation matrices, significantly reducing trainable parameters while preserving model's expressiveness.
- 2) Prompt-based Tuning (Lester et al., 2021): This approach optimizes continuous or discrete prompts in the input space to steer the model's behavior toward specific tasks. By modifying only the prompt embeddings while keeping model weight frozen, it achieves remarkable parameter efficiency in task adaptation.
- 3) Adapter-based Tuning (Houlsby et al., 2019): This strategy incorporates specialized adapter modules between the layers of the pre-trained model. By updating only these compact adapters while freezing the original model parameters, it enables efficient task-specific adaptation with minimal architectural modifications to the base model.
- 4) Selective-based Tuning (Zaken et al., 2022): This approach focuses on selectively fine-tuning specific layers or parameters of the model that are most relevant to the downstream task. Through careful selection, it significantly reduces the resource consumption.
- 5) Other Tuning Methods (Li et al., 2024e): This category encompasses techniques like zeroth-order optimization (Malladi et al., 2023), split learning (Thapa et al., 2022), model compression (Deng et al., 2020), and data selection (Qin et al., 2024), which offer innovative ways to optimize LLM performance with lower resource requirements.

3 Challenges

In this section, we provide an in-depth analysis of the challenges encountered in FedLLM, focusing on four key aspects: communication overhead, data heterogeneity, memory constraints, and computation burden.

3.1 Communication Overhead

In federated fine-tuning, the learning process necessitates iterative communication between participating clients and the central server, where clients periodically transmit their locally updated model parameters for aggregation (Fu et al., 2022; 2024b;c). This iterative exchange continues until model convergence, inherently introducing substantial communication overhead (Li et al., 2022b; Kou et al., 2025). The challenge is even more pronounced when fine-tuning LLMs, which consist of billions of parameters. To quantify this challenge, Figure 5 presents a comparative analysis of parameter sizes across different models, contrasting traditional models like BERT (Devlin et al., 2019b) with the LLaMA series, including TinyLLaMA (Zhang et al., 2024d), LLaMA2-7B, LLaMA2-13B (Touvron et al., 2023b), LLaMA3-3B, and LLaMA3-8B (Dubey et al., 2024). Our analysis reveals that LLaMA models are dramatically larger than BERT, with parameter counts ranging from 10 to 118× greater. This exponential increase in parameter size directly translates to significantly higher data transmission volumes in each communication round, substantially elevating bandwidth requirements and overall communication costs in federated environments.

However, in real-world scenarios, communication bandwidth is often severely constrained. According to a 2023 Cisco report, approximately 30% of edge devices still rely on 2G or 3G networks, which provide bandwidths of less than 10 Mb/s (Wang et al., 2023c). While 5G networks offer speeds more than 50× faster, they are accessible to only about 10% of devices. This significant disparity in network capabilities inevitably results in substantial communication delays, particularly when exchanging large parameter updates. More critically, the duration of each training round is determined by the slowest device in the network—a phenomenon known as the "straggler effect." This means that devices with limited connectivity can dramatically hinder the convergence speed of the federated fine-tuning process. Consequently, minimizing communication overhead becomes essential for effective FedLLM implementation. Efficient management of data transmission can accelerate model convergence while ensuring the practical feasibility of deploying FedLLM in bandwidth-constrained environments. Without addressing the communication challenge, the theoretical privacy benefits of federated fine-tuning may remain inaccessible to many real-world applications, particularly those involving edge devices or regions with limited network infrastructure.

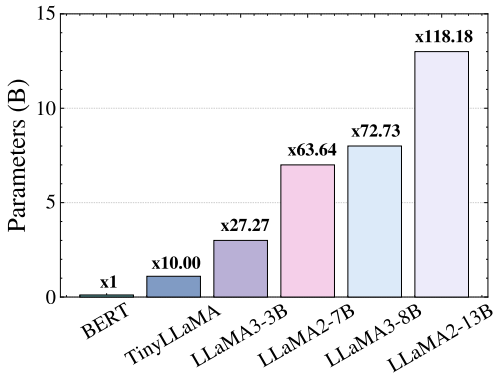

Figure 5: Comparison of model parameters across BERT and LLaMA series models.

3.2 Data Heterogeneity

Data heterogeneity is a notorious challenge in FL, manifesting in significant variations in data distribution (Tian et al., 2022a), quality (Tam et al., 2023a;b), and quantity (Yi et al., 2022) across clients. Such heterogeneity hinders convergence and degrades the global model's generalization ability, as it must reconcile conflicting updates derived from diverse client populations (Ma et al., 2024a; Wang et al., 2025). To mitigate the adverse effects of data heterogeneity, various strategies have been explored in traditional FL, which can be broadly categorized into four groups: 1) **Regularization-based methods** incorporate additional penalty terms into the local objective to limit model divergence and encourage alignment with the global model (Li et al., 2020); 2) **Aggregation-based methods** modify the server-side aggregation strategy to assign adaptive weights to client updates, reducing the influence of noisy, unreliable, or biased data sources (Wang et al., 2020a;b); 3) **Data-sharing methods** introduce small, carefully curated auxiliary datasets that are distributed to clients to promote distributional alignment and reduce inter-client drift (Goetz & Tewari, 2020); 4) **Personalized FL approaches** aim to learn client-specific models that better capture the unique characteristics of local data (Kou et al., 2024). While these techniques have demonstrated success in traditional FL scenarios, addressing data heterogeneity in the context of FedLLM remains largely underexplored. This challenge is further exacerbated when applying PEFT techniques, as PEFT tends to be more sensitive to distributional shifts and limited data availability. As illustrated in Figure 6, the performance gap between PEFT and full-parameter fine-tuning (FFT) grows wider as data heterogeneity increases, underscoring the need for targeted solutions to improve PEFT robustness in federated settings.

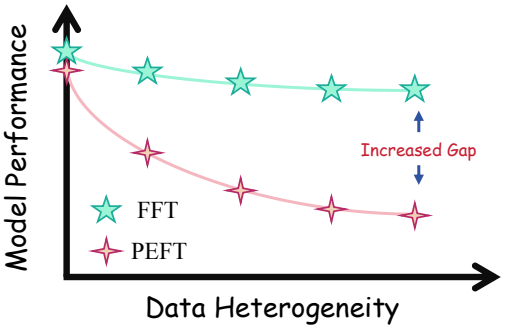

Figure 6: Impact of data heterogeneity on model performance. As the degree of data heterogeneity increases, the performance gap between PEFT and FFT widens.

3.3 Memory Wall

Memory constraints present a fundamental challenge to the practical deployment of federated fine-tuning (Wu et al., 2024j; Wu et al.). During the local fine-tuning process, model parameters, intermediate activations, and gradients must be stored in memory, resulting in substantial memory consumption. However, participating

clients, especially edge devices, typically have limited available memory, ranging from 4 to 12 GB (Tian et al., 2024a). This limited memory capacity is insufficient to support fine-tuning mainstream LLMs.

To better quantify this challenge, we profile the memory usage during full-parameter fine-tuning for both traditional models (e.g., DistilBERT (Sanh et al., 2020), BERT (Devlin et al., 2019a)) and LLaMA-series models (e.g., TinyLLaMA, LLaMA2-7B, LLaMA2-13B). As shown in Figure 7, our results reveal a dramatic disparity in resource requirements between these model families. Fine-tuning LLaMA models demands substantially higher memory resources compared to traditional architectures. Specifically, fine-tuning LLaMA2-7B requires approximately 51.85 GB of GPU memory, which is 7.68× more than BERT (6.75 GB). This requirement escalates further with LLaMA2-13B, which demands 98.56 GB of memory, representing a 28.32× increase over DistilBERT and vastly exceeding the available memory capacity of edge devices. This stark mismatch between the memory demands of fine-tuning LLMs and the hardware limitations of participants creates a **Memory Wall**, a fundamental barrier that severely restricts the feasibility of deploying FedLLM at scale. This memory constraint prevents a significant proportion of devices from participating in the collaborative learning process, thereby compromising model performance through reduced data diversity and limiting the practical application scope of FedLLM. The memory wall represents not just a technical challenge but a fundamental constraint on democratizing access to advanced AI capabilities through FL.

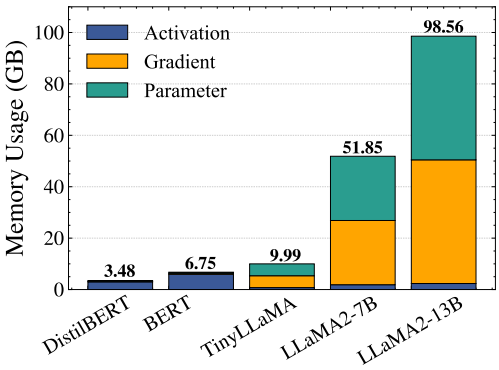

Figure 7: Memory usage and breakdown analysis when fine-tuning different models.

3.4 Computation Overhead

Computational cost presents another major bottleneck in deploying FedLLM (Tian et al., 2022a; Almanifi et al., 2023). The computational demands of fine-tuning LLMs arise from the forward and backward passes during local training iterations, each contributing significantly to the overall processing burden. The sheer scale of these models—characterized by billions of parameters, numerous transformer layers, and complex attention mechanisms—makes them inherently compute-intensive, which can quickly overwhelm devices with limited processing capabilities. Moreover, the iterative nature of fine-tuning, which involves repeated forward and backward passes, compounds the computational load, making it difficult to achieve efficient training on resource-constrained devices. To quantitatively understand this challenge, we profile the computational demands of fine-tuning various models by measuring the floating-point operations (FLOPs) required for a single forward and backward pass with a batch size of 16.

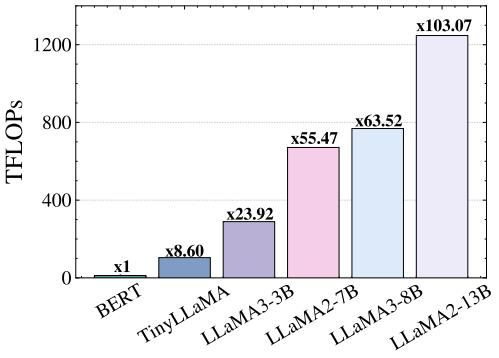

Figure 8: Comparison of FLOPs for a single forward and backward pass across models.

Specifically, we evaluate BERT alongside a suite of LLaMA-based models, including TinyLLaMA, LLaMA3-3B, LLaMA2-7B, LLaMA3-8B, and LLaMA2-13B. As shown in Figure 8, our results reveal a dramatic escalation in computational requirements for LLaMA-series models compared to BERT. For instance, fine-tuning TinyLLaMA incurs 8.60× FLOPs of BERT, while LLaMA2-13B demands a staggering 103.07× more FLOPs. This exponential increase in computational complexity directly results in significantly longer training time, excessive energy consumption on battery-powered devices, and thermal management issues, all of which can degrade hardware performance over time, thereby undermining the feasibility of large-scale, real-world deployment (Ning et al., 2024; Tian et al., 2023). These findings highlight the pressing need for computation-efficient fine-tuning strategies that can effectively accommodate the heterogeneous and resource-constrained nature of participating devices, while ensuring model performance.

4 Federated Fine-tuning

In this section, we introduce various parameter-efficient fine-tuning methods and discuss their applications in FL. Figure 9 provides an overview of representative parameter-efficient federated fine-tuning methods.

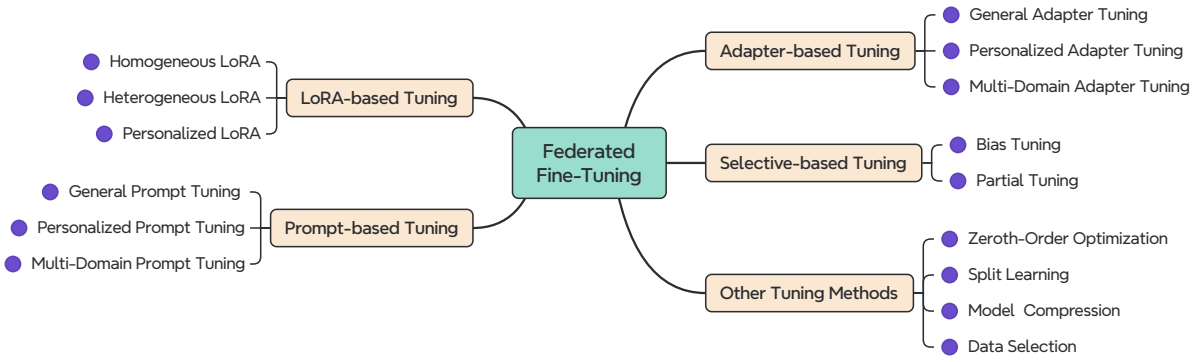

Figure 9: Overview of parameter-efficient federated fine-tuning methods and their corresponding taxonomy.

4.1 LoRA-based Tuning

4.1.1 Preliminary

Low-Rank Adaptation (LoRA) (Hu et al., 2021; Tian et al., 2024c) has emerged as a promising approach for efficient fine-tuning of LLMs while maintaining model performance. The core idea of LoRA lies in introducing low-rank matrices into the pre-trained model's weights, allowing for the adaptation of model parameters without altering the original architecture significantly. LoRA is based on the observation that fine-tuning does not require updating the full parameter space; instead, meaningful adaptations can often be represented in a low-dimensional subspace. By applying low-rank decomposition to the weight updates, LoRA drastically reduces the number of trainable parameters, leading to lower resource consumption. Furthermore, LoRA's modular design allows it to be easily integrated into a variety of model architectures without altering the original model.

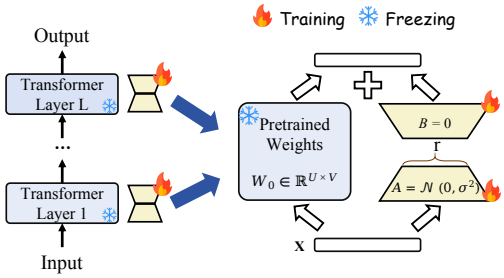

Figure 10: The working principle of LoRA.

Figure 10 illustrates the working principle of LoRA. Specifically, the pre-trained parameter matrix $\mathbf{W_0} \in \mathbb{R}^{U \times V}$ is decomposed into two matrices, $\mathbf{A} \in \mathbb{R}^{r \times V}$ and $\mathbf{B} \in \mathbb{R}^{U \times r}$, where $r \ll \min(U, V)$ denotes the rank controlling the dimensionality of the low-rank subspace. The matrix $\mathbf{A}$ projects the input into a low-dimensional space, and the matrix $\mathbf{B}$ maps it back to the original space. During fine-tuning, only $\mathbf{A}$ and $\mathbf{B}$ are updated, while the original weights $\mathbf{W}_0$ remain frozen. The input data X is processed by both $\mathbf{W_0}$ and $\mathbf{BA}$. The output of $\mathbf{W_0}X$ is the initial prediction generated by the pre-trained model, while the output of $\mathbf{BA}X$ represents the task-specific adaptation introduced by the low-rank matrices. These two outputs are then added element-wise to produce the final output. This process can be formulated as:

$$h = \mathbf{W_0}X + \mathbf{BA}X \tag{3}$$

By updating only the low-rank matrices $\mathbf{A}$ and $\mathbf{B}$, LoRA enables efficient task adaptation with minimal computational overhead, effectively capturing task-specific knowledge while preserving the generalization ability of the original pre-trained model.

4.1.2 LoRA in Federated Fine-tuning

In the context of federated fine-tuning, LoRA offers notable advantages in both communication and computation efficiency. Since only the low-rank matrices are updated and transmitted, rather than the full model parameters, this lightweight updating mechanism significantly lowers bandwidth requirements and facilitates

Table 2: **Homogeneous LoRA in federated fine-tuning.**

Method	Challenge			
	Communication	Non-IID	Memory	Computation
FedIT (Zhang et al., 2024b)	✗	✗	✗	✗
FedSA-LoRA (Guo et al., 2024b)	✓	✗	✗	✗
FederatedScope-LLM (Kuang et al., 2023)	✓	✗	✗	✓
FeDeRA (Yan et al., 2024)	✗	✓	✗	✗
LoRA-FAIR (Bian et al., 2024)	✗	✗	✗	✗
FLASC (Kuo et al., 2024)	✓	✗	✗	✗
SA-FedLora (Yang et al., 2024c)	✓	✓	✗	✗
SLoRA (Babakniya et al., 2023)	✗	✓	✗	✗
RoLoRA (Chen et al., 2024d)	✗	✓	✗	✗
FedPipe (Fang et al., 2024b)	✓	✗	✓	✓
Lp-FL (Jiang et al., 2023)	✗	✗	✗	✗
Fed-piLot (Zhang et al., 2024l)	✗	✓	✓	✗
FedRA (Su et al., 2023)	✓	✗	✓	✓

faster convergence of the global model, making LoRA particularly well-suited for federated environments. In this paper, we introduce a novel taxonomy of LoRA-based federated fine-tuning methods, categorizing them into three primary types: Homogeneous LoRA, where all clients adopt the same rank; Heterogeneous LoRA, where clients use different ranks based on their resources; and Personalized LoRA, which tailors low-rank adaptations to individual client data distributions. This taxonomy is summarized in Figure 9. In the following sections, we delve into representative methods within each category, analyzing how they address the key challenges identified in Section 3, beyond the inherent benefits brought by LoRA itself.

• **Homogeneous LoRA** refers to scenarios where all clients adopt the same low-rank dimension r for their LoRA modules. This uniform configuration simplifies aggregation and model synchronization across clients. Table 2 summarizes representative methods in this category and the specific challenges they address.

FedIT (Zhang et al., 2024b) directly integrates LoRA into the classic FedAvg (McMahan et al., 2017) for instruction tuning. **FedSA-LoRA** (Guo et al., 2024b) identifies that the **A** matrices primarily encode general knowledge, while the **B** matrices capture client-specific features; thus, it only uploads **A** to the server, significantly reducing communication overhead. **FederatedScope-LLM** (Kuang et al., 2023) establishes a comprehensive end-to-end pipeline for federated LLM fine-tuning and proposes offsite-tuning strategies to mitigate both communication and computational costs. **FeDeRA** (Yan et al., 2024) addresses data heterogeneity by initializing LoRA matrices via singular value decomposition on the pre-trained weights. **LoRA-FAIR** (Bian et al., 2024) introduces a correction mechanism on the server to handle aggregation bias and initialization drift across clients. **FLASC** (Kuo et al., 2024) incorporates sparsity into LoRA to further reduce communication overhead.

SA-FedLoRA (Yang et al., 2024c) mitigates client drift through parameter regularization and dynamically allocates communication budgets. **SLoRA** (Babakniya et al., 2023) proposes a novel data-driven initialization scheme to better handle statistical heterogeneity. **RoLoRA** (Chen et al., 2024d) adopts an alternating minimization approach to improve robustness under non-IID conditions. **FedPipe** (Fang et al., 2024b) automatically selects critical parameters for fine-tuning and applies quantization to reduce memory usage. **LP-FL** (Jiang et al., 2023) applies LoRA directly to enable efficient on-device fine-tuning. **Fed-piLot** (Zhang et al., 2024l) reduces memory consumption through LoRA assignment strategies and introduces a novel spatial-temporal aggregation (STAgg) rule to address heterogeneity. **FedRA** (Su et al., 2023) adaptively determines parameter update scopes based on client resource constraints, effectively reducing computational, communication, and memory costs.

• **Heterogeneous LoRA** allows individual clients to adopt different rank values r for their LoRA modules, based on their specific data characteristics or resource constraints. This heterogeneity can manifest either across clients (inter-model) or within different layers of the same model (intra-model). By enabling each client to select a rank that best fits its capabilities and local data, this approach introduces greater flexibility and resource-awareness into the federated fine-tuning process. Table 3 summarizes representative methods and the specific challenges they address.

Table 3: **Heterogeneous LoRA in federated fine-tuning.**

Method	Challenge			
	Communication	Non-IID	Memory	Computation
HETLoRA (Cho et al., 2024)	✓	✓	✓	✓
FLoRA (Wang et al., 2024f)	✓	✗	✓	✓
FlexLoRA (Bai et al., 2024a)	✓	✗	✓	✓
LoRA-A^2 (Koo et al., 2024)	✓	✓	✓	✓
Byun & Lee (2024)	✓	✗	✓	✓
FedHM (Yao et al., 2021)	✓	✗	✓	✓
RBLA (Tavallaie & Nazemi[1])	✓	✓	✓	✓

Table 4: **Personalized LoRA in federated fine-tuning.**

Method	Challenge			
	Communication	Non-IID	Memory	Computation
FDLoRA (Qi et al., 2024a)	✗	✓	✗	✗
pFedLoRA (Yi et al., 2023)	✗	✓	✗	✗
FedLoRA (Wu et al., 2024f)	✗	✓	✗	✗
FedDPA (Yang et al., 2024b)	✗	✓	✗	✗
PerFIT (Zhang et al., 2024e)	✗	✓	✗	✗
FedMEM (Du et al., 2024)	✗	✓	✗	✗
FedAMoLE (Zhang et al., 2024i)	✗	✓	✗	✗

HETLoRA (Cho et al., 2024) assigns heterogeneous ranks across devices and incorporates rank self-pruning along with sparsity-weighted aggregation to tackle data heterogeneity. **FLoRA** (Wang et al., 2024f) proposes a stacking-based aggregation scheme and allows devices to select ranks according to their resource budgets. **FlexLoRA** (Bai et al., 2024a) enables dynamic adjustment of local LoRA ranks to leverage the heterogeneous device resources, while employing singular value decomposition for weight redistribution. **LoRA-A^2** (Koo et al., 2024) introduces alternating freezing and adaptive rank selection mechanisms to fully utilize heterogeneous device resources while addressing statistical heterogeneity. Byun & Lee (2024) propose a replication-based padding technique to enable aggregation across clients with varying LoRA ranks. **FEDHM** (Yao et al., 2021) addresses resource constraints by distributing low-rank models with heterogeneous capacities to clients. **RBLA** (Tavallaie & Nazemi[1]) improves aggregation robustness by simultaneously maintaining and aligning both low-rank and high-rank feature components.

• **Personalized LoRA** enables each participant to fine-tune its model using personalized low-rank adaptation matrices, allowing for better alignment with local data characteristics. This approach enhances the ability of the global model to generalize across clients while retaining client-specific nuances. Table 4 summarizes representative methods and the specific challenges they aim to address.

FDLoRA (Qi et al., 2024a) introduces dual LoRA modules on each client to separately capture global and personalized knowledge. Only the global LoRA module are communicated with the central server for aggregating cross-client knowledge. **pFedLoRA** (Yi et al., 2023) designs a homogeneous small adapter to facilitate federated clients' heterogeneous local model training, with a proposed iterative training process for global-local knowledge exchange. **FedLoRA** (Wu et al., 2024f) maintains shared general knowledge in a global full-rank matrix while encoding client-specific knowledge in a personalized low-rank module. **FedDPA** (Yang et al., 2024b) utilizes a global adapter and a local adapter to jointly address test-time distribution shifts and client-specific personalization. **PerFIT** (Zhang et al., 2024e) allows each client to search for a personalized architecture by expanding the trainable parameter space of the global model to address data heterogeneity. **FEDMEM** (Du et al., 2024) equips the global model with a k-nearest neighbor (KNN) classifier that captures client-specific distributional shifts, achieving personalization and overcoming data heterogeneity. **FedAMoLE** (Zhang et al., 2024i) features an adaptive mixture of LoRA experts (MoLE) module for aggregating heterogeneous models and a reverse selection-based expert assignment strategy that optimizes model architectures based on data distributions.

4.2 Prompt-based Tuning

4.2.1 Preliminary

Prompt-based tuning (Lester et al., 2021) has emerged as a highly effective and resource-efficient alternative to conventional fine-tuning approaches for LLMs. Unlike traditional methods that update the model's parameters directly, prompt-based tuning learns a set of trainable prompts or input embeddings that steer the model's behavior on downstream tasks. By modifying only the input space, this approach leverages the pre-trained knowledge of LLMs without altering their weights. As illustrated in Figure 11, a sequence of trainable prompt embeddings $P \in \mathbb{R}^{l_p \times d}$ is prepended to the original input tokens $X \in \mathbb{R}^{l_x \times d}$, where l_p and l_x denote the lengths of the prompt and input sequences, respectively, and d represents the model's hidden dimension. The concatenated sequence is then fed into the frozen model: $Z = f([P; X]; \theta)$, where $f(\cdot; \theta)$ denotes the pre-trained LLM with frozen parameters θ, and $[P; X]$ represents the concatenation of the trainable prompts and the original input tokens. By optimizing the prompt embeddings P, the model can effectively adapt to new tasks while reusing its pre-trained knowledge. This approach enables efficient task adaptation without modifying the model weights, thereby significantly reducing memory and computational overhead. Through prompt-based tuning, task-specific guidance is embedded within the prompts, allowing the model to generate desired outputs by attending to relevant information stored in its pre-trained parameters.

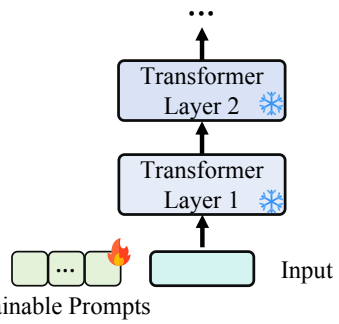

Figure 11: The working principle of prompt tuning.

4.2.2 Prompt in Federated Fine-tuning

In the context of federated fine-tuning, prompt-based tuning offers significant advantages in communication efficiency and model adaptability. Since only the trainable prompt embeddings are updated and exchanged, rather than model parameters, this approach substantially reduces communication overhead between clients and the central server. Additionally, by freezing the base model, prompt-based tuning allows clients with heterogeneous data distributions to personalize their behavior effectively, while still benefiting from globally shared knowledge encoded in the pre-trained model. These properties make prompt-based tuning a compelling choice for federated fine-tuning. In this paper, we propose a novel taxonomy of prompt-based federated fine-tuning approaches, categorizing them into three primary types: General Prompt Tuning, Personalized Prompt Tuning, and Multi-Domain Prompt Tuning, as illustrated in Figure 9. In the following sections, we examine representative methods within each category and analyze how they address the challenges outlined in Section 3, beyond the inherent benefits brought by prompt tuning itself.

• **General Prompt Tuning** refers to approaches in which a shared set of prompt embeddings is learned and applied uniformly across all participating clients. In this setting, the same prompts are prepended to each client's input sequences, providing consistent task-specific guidance and enabling the global model to generalize across diverse data sources. Table 5 summarizes representative methods and the specific challenges they aim to address.

MetePFL (Chen et al., 2023a) applies prompt tuning to fine-tune a spatio-temporal Transformer-based foundation model for weather forecasting tasks in a federated setting. **PromptFL** (Guo et al., 2023c) adapts CLIP models for vision-language tasks in FL using prompt-based tuning. **FedBPT** (Sun et al., 2023) employs prompt-based tuning to efficiently adapt black-box LLMs using gradient-free optimization, eliminating the need for clients to access model parameters and requiring only forward propagation for local training. **FedPepTAO** (Che et al., 2023) introduces a partial prompt tuning mechanism to reduce communication costs, along with an adaptive optimization algorithm to address data heterogeneity. **FedBBPT** (Lin et al., 2023) enables clients to utilize a zeroth-order optimizer locally, obviating the need for full LLM deployment, effectively reducing memory consumption and computational costs. **FedTPG** (Qiu et al., 2023) learns a unified, task-aware prompt generation network conditioned on input text, improving generalization to both seen and unseen classes.

Table 5: **General prompt tuning in federated fine-tuning.**

Method	Challenge			
	Communication	Non-IID	Memory	Computation
MetePFL (Chen et al., 2023a)	✗	✗	✗	✗
PromptFL (Guo et al., 2023c)	✗	✗	✗	✗
FedBPT (Sun et al., 2023)	✗	✗	✓	✓
FedPepTAO (Che et al., 2023)	✓	✓	✗	✗
Fed-BBPT (Lin et al., 2023)	✗	✗	✓	✓
FedTPG (Qiu et al., 2023)	✗	✗	✗	✗
FedPR (Feng et al., 2023a)	✗	✓	✗	✗
Fed-CPrompt (Bagwe et al., 2023)	✗	✓	✗	✗
Fedprompt (Zhao et al., 2023a)	✗	✗	✗	✗
FedSP (Dong et al., 2023a)	✗	✗	✓	✓
HePCo (Halbe et al., 2023)	✗	✓	✗	✗
PFL-GCN (Ahmad et al., 2023)	✗	✗	✗	✗
AUG-FedPrompt (Cai et al., 2023b)	✗	✗	✗	✗
Liu et al. (2023e)	✗	✗	✗	✗
FedHPL (Ma et al., 2024b)	✗	✓	✗	✗
PFPT (Weng et al.)	✗	✓	✗	✗
FCILPT (Liu et al., 2023c)	✗	✓	✗	✗
CaFPT (Guo et al., 2024c)	✗	✗	✗	✗
FedPoD (Chen et al., 2023b)	✗	✓	✗	✗

FedPR (Feng et al., 2023a) enhances federated visual prompt tuning by projecting local prompt updates into an approximate null space of the global prompt, mitigating gradient interference and improving global performance. **Fed-CPrompt** (Bagwe et al., 2023) addresses asynchronous task arrivals and heterogeneous data distributions via asynchronous prompt updates and a contrastive continual learning loss. **FedPrompt** (Zhao et al., 2023a) employs a split aggregation strategy, freezing the extensive parameters of LLMs and only tuning and aggregating soft prompts. **FedSP** (Dong et al., 2023a) reduces computational and memory overhead by utilizing a lightweight auxiliary model for prompt learning. **HePCo** (Halbe et al., 2023) mitigates catastrophic forgetting and data heterogeneity through a data-free distillation method performed in the model's latent space. **PFL-GCN** (Ahmad et al., 2023) employs prompt tuning specifically for sentiment analysis.

AUG-FedPrompt (Cai et al., 2023b) exploits abundant unlabeled data for data augmentation to address the issue of data scarcity. Liu et al. (2023e) integrate self-consistency and chain-of-thought prompting to improve zero-shot performance of LLMs. **FedHPL** (Ma et al., 2024b) introduces a global logit distillation framework to handle model heterogeneity and guide the local training process. **PFPT** (Weng et al.) proposes a probabilistic prompt aggregation mechanism to address data heterogeneity and imbalanced data distribution. **FCILPT** (Liu et al., 2023c) jointly encodes task-relevant and task-irrelevant knowledge into prompts to preserve both previous and newly learned knowledge, alleviating catastrophic forgetting. **CaFPT** (Guo et al., 2024c) leverages information-theoretic principles to facilitates the retrieval process by conditioning on examples that activate the most relevant knowledge inside pre-trained models. **FedPoD** (Chen et al., 2023b) employs lightweight prompts to guide frozen foundation models and introduces multi-level prompt-based communication to enable multi-source knowledge fusion and controlled optimization.

• **Personalized Prompt Tuning** enables each client to tailor its prompt embeddings based on local data distributions and task-specific requirements. By fine-tuning prompts individually, clients can better capture local nuances and context-specific information that a one-size-fits-all prompt might overlook. This approach directly addresses the challenge of data heterogeneity by facilitating local adaptation, while still allowing clients to benefit from global knowledge aggregated during training. Table 6 summarizes representative methods and the specific challenges they target.

pFedPG (Yang et al., 2023a) deploys a personalized prompt generator on the server to produce client-specific visual prompts, enabling efficient adaptation of frozen backbones to diverse local data. **SGPT** (Deng et al., 2024) combines generalized and personalized FL by learning a mix of shared and group-specific prompts to capture both commonalities and group-specific variations. **pFedPrompt** (Guo et al., 2023b) leverages the unique multimodal capabilities of vision-language models by learning client consensus in the linguistic space and adapting to client characteristics in the visual space in a non-parametric manner. **FedOTP** (Li et al.,

Table 6: **Personalized prompt tuning in federated fine-tuning.**

Method	Challenge			
	Communication	Non-IID	Memory	Computation
pFedPG (Yang et al., 2023a)	✗	✓	✗	✗
SGPT (Deng et al., 2024)	✗	✓	✗	✗
pFedPrompt (Guo et al., 2023b)	✗	✓	✗	✗
FedOTP (Li et al., 2024c)	✗	✓	✗	✗
pFedPT (Li et al., 2023b)	✗	✓	✗	✗
FedMGP (Yu et al., 2024)	✗	✓	✗	✗
FedLPPA (Lin et al., 2024)	✗	✓	✗	✗
FedPGP (Cui et al., 2024)	✗	✓	✗	✗
FedPFT (Wu et al., 2024g)	✗	✓	✗	✗
Wang et al. (2024c)	✓	✓	✓	✓
pFedMoAP (Luo et al., 2024)	✗	✓	✗	✗
CP^{2}GFed (Gao et al., 2024a)	✗	✓	✗	✓

Table 7: **Multi-domain prompt tuning in federated fine-tuning.**

Method	Challenge			
	Communication	Non-IID	Memory	Computation
DiPrompT (Bai et al., 2024b)	✗	✓	✗	✗
PFCR (Guo et al., 2024a)	✗	✓	✗	✗
Fed-DPT (Wei et al., 2023)	✗	✓	✗	✗
FedAPT (Su et al., 2024)	✗	✓	✗	✗
Zhao et al. (2024b)	✗	✓	✗	✗
FedDG (Gong et al., 2024)	✗	✓	✗	✗
CP-Prompt (Feng et al., 2024b)	✗	✓	✗	✗

2024c) introduces efficient collaborative prompt learning strategies to capture diverse category traits on a per-client basis. **pFedPT** (Li et al., 2023b) utilizes personalized visual prompts to implicitly represent local data distribution information and provides this information to the aggregation model to enhance classification tasks. **FedMGP** (Yu et al., 2024) uses coarse-grained global prompts for shared knowledge and fine-grained local prompts for personalization, and introduces a selective fusion mechanism for prompt aggregation.

FedLPPA (Lin et al., 2024) jointly learns personalized prompts and aggregation strategies for weakly-supervised medical image segmentation. **FedPGP** (Cui et al., 2024) employs pre-trained CLIP to provide knowledge-guidance for the global prompt, enhancing generalization while incorporating a low-rank adaptation term to personalize the global prompt. **FedPFT** (Wu et al., 2024g) addresses feature-classifier mismatch through prompt-driven feature transformation. Wang et al. (2024c) propose a discrete local search strategy for gradient-free local training and a token-based compression method inspired by linear word analogies, substantially reducing resource costs. **pFedMoAP** (Luo et al., 2024) introduces a personalized prompt learning framework based on the mixture-of-experts paradigm (Cai et al., 2024a). **CP^{2}GFed** (Gao et al., 2024a) introduces a cross-granularity knowledge transfer mechanism and dynamic personalized prompt generation to improve model performance.

• **Multi-Domain Prompt Tuning** extends the prompt-based approach to environments where federated clients operate across distinct domains or application contexts. In such scenarios, each client is equipped with domain-specific prompt embeddings that adapt the shared global model to diverse contextual and distributional conditions. This approach enhances the model's generalization ability across heterogeneous domains while maintaining a shared global foundation. It is particularly valuable in real-world deployments spanning multiple industries or task categories. Table 7 summarizes representative methods and the specific challenges they address.

DiPrompT (Bai et al., 2024b) proposes a distributed domain generalization approach using adaptive prompts, introducing global prompts for shared knowledge and domain prompts for domain-specific adaptation. **PFCR** (Guo et al., 2024a) eliminates the need for raw data sharing via encrypted gradient updates, models items in a unified feature space using descriptive text, and facilitates cross-domain knowledge transfer

through federated content representations and prompt tuning. **Fed-DPT** (Wei et al., 2023) leverages a pre-trained vision-language model and applies dual prompt tuning—combining visual and textual prompts—for improved domain alignment across decentralized data sources. **FedAPT** (Su et al., 2024) introduces a meta prompt, an adaptive network, and frozen keys to personalize prompts for each test sample, thereby enhancing multi-domain image classification. Zhao et al. (2024b) propose a language distance metric to improve data efficiency and facilitate cross-linguistic generalization. **FedDG** (Gong et al., 2024) allows clients to learn text and visual prompts locally while maintaining indirect alignment via global prompts used as a shared reference. Domain-specific prompts are exchanged among clients and selectively integrated into global prompts using lightweight attention-based aggregators. **CP-Prompt** (Feng et al., 2024b) captures intra-domain knowledge by inserting personalized prompts into the multi-head attention modules and subsequently learns inter-domain representations through a shared prompting mechanism.

4.3 Adapter-based Tuning

4.3.1 Preliminary

Adapter-based tuning is another parameter-efficient alternative to full-parameter fine-tuning for LLMs (Pfeiffer et al., 2020). It introduces lightweight, trainable adapter modules into the model while keeping the pre-trained weights frozen. These modules act as task-specific components that transform intermediate representations in a controlled manner, enabling efficient adaptation to downstream tasks with minimal memory and computational overhead. A standard adapter module consists of three key operations: **down-projection**, **non-linearity**, and **up-projection**, as shown in Figure 12. For activations $h_i \in \mathbb{R}^{n \times d}$, the adapter transformation proceeds as follows:

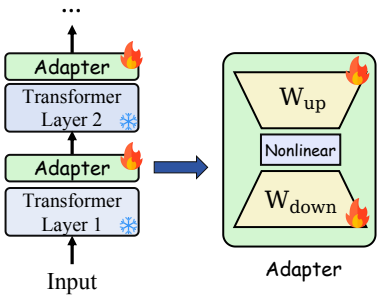

Figure 12: The working principle of adapter tuning.

1) Down-Projection: The high-dimensional hidden state h_i is projected into a low-dimensional space using a learnable weight matrix $W_{DP} \in \mathbb{R}^{d \times r}$, where $r \ll d$ controls the bottleneck size. This step reduces the number of trainable parameters while capturing essential features:

$$h_i^{'} = h_i W_{DP} \tag{4}$$

2) Non-Linearity: A non-linear activation function $\sigma(\cdot)$, such as ReLU or GELU, is applied to introduce expressive transformations while retaining important task-specific patterns:

$$h_i^{''} = \sigma(h_i^{'}) \tag{5}$$

3) Up-Projection: The transformed low-dimensional representation is mapped back to the original feature space using an **up-projection** matrix $W_{UP} \in \mathbb{R}^{r \times d}$:

$$h_i^{'''} = h_i^{''} W_{UP} \tag{6}$$

4) Residual Connection: The final output of the adapter is then residually added to the original hidden state, preserving the pre-trained knowledge while incorporating task-specific adjustments:

$$Z = h_i + h_i^{'''} \tag{7}$$

where Z represents the adapted hidden representation. This residual connection ensures that the pre-trained model remains largely intact while allowing task-specific fine-tuning through the lightweight adapter layers. Notably, only the adapter parameters W_{DP} and W_{UP} are updated during training, resulting in significantly lower memory and computation costs compared to full-parameter fine-tuning.

4.3.2 Adapter in Federated Fine-tuning

In the context of federated fine-tuning, adapter-based tuning provides significant advantages in both resource efficiency and model adaptability. Since only the lightweight adapter modules are updated and exchanged,

Table 8: **General, personalized, and multi-domain adapter tuning in federated fine-tuning.**

Method	Type	Challenge			
		Communication	Non-IID	Memory	Computation
FedAdapter (Cai et al., 2023a)	General	✓	✗	✗	✓
Kim et al. (2023a)	General	✓	✓	✓	✓
FedTT+ (Ghiasvand et al., 2024)	General	✓	✓	✗	✗
C2A (Kim et al., 2023b)	Personalized	✗	✓	✗	✗
FedCLIP (Lu et al., 2023b)	Personalized	✗	✓	✗	✗
Fed-MNMT (Liu et al., 2023f)	Multi-domain	✗	✓	✗	✗
AdaFedSelecKD (Feng et al., 2024a)	Multi-domain	✗	✓	✗	✗
FedDAT (Chen et al., 2024a)	Multi-domain	✗	✓	✗	✗

rather than the full model parameters, this approach greatly reduces computation and communication overhead. Moreover, by freezing the base model, adapter-based tuning enables clients to fine-tune efficiently on heterogeneous local data while still benefiting from the shared global knowledge encoded in the pre-trained model. This modular design facilitates the seamless integration of task-specific adaptations without compromising the generalization capability of the base model. To better understand the landscape of adapter-based methods in federated fine-tuning, we propose a new taxonomy comprising three categories: General Adapter Tuning, Personalized Adapter Tuning, and Multi-Domain Adapter Tuning, as illustrated in Figure 9. In the following sections, we explore representative methods within each category and analyze how they address the core challenges identified in Section 3, beyond the inherent benefits brought by adapter itself.

• **General Adapter Tuning** refers to scenarios in which all clients utilize a shared adapter structure with identical initialization. In this setting, the same adapter modules are inserted into the transformer layers of each client's model, enabling consistent adaptation mechanisms across the federation. This uniformity facilitates stable aggregation and coordinated updates during federated training. Such an approach is particularly effective when clients operate on similar tasks or share relatively homogeneous data distributions, as a globally optimized adapter can generalize well across participants. Table 8 summarizes representative methods and the specific challenges they aim to address.

FedAdapter (Cai et al., 2023a) proposes a progressive adapter tuning strategy, combined with continuous device profiling, to dynamically optimize adapter configurations across clients, improving efficiency without sacrificing accuracy. Kim et al. (2023a) leverage adapters to address the high communication costs associated with federated fine-tuning of LLMs. **FedTT+** (Ghiasvand et al., 2024) integrates tensorized adapters for LLM adaptation and further improves robustness to data heterogeneity by freezing portions of the tensor factors, significantly reducing the number of trainable parameters while maintaining model performance.

• **Personalized Adapter Tuning** enables each client to independently fine-tune its adapter modules based on its local data distribution and task-specific requirements. In contrast to general adapter tuning, this approach does not enforce uniformity across clients; instead, it allows for the retention of personalized adapter parameters that better capture client-specific knowledge. This strategy is particularly advantageous in federated settings characterized by high degrees of data heterogeneity. By leveraging personalized adapters, clients can achieve improved local performance while still benefiting from shared global knowledge. Table 8 summarizes representative methods and the specific challenges they address.

C2A (Kim et al., 2023b) employs a hypernetwork to generate client-specific adapters, effectively addressing data heterogeneity by enabling on-demand parameter generation tailored to each client. **FedCLIP** (Lu et al., 2023b) introduces an attention-based adapter design that utilizes the pre-trained model's knowledge to facilitate both rapid generalization and efficient personalization while minimizing resource overhead.

• **Multi-Domain Adapter Tuning** extends the federated fine-tuning paradigm to clients operating across distinct domains, enabling efficient adaptation to domain-specific tasks. In this setting, each client maintains its own domain-specific adapter while contributing to a shared global model. The global model aggregates adapter updates across domains to capture domain-invariant representations to support generalization. This approach is particularly effective in cross-domain scenarios such as multilingual natural language processing. By decoupling domain-specific learning from the shared backbone, this strategy balances personalization and collaboration. Table 8 summarizes representative methods and the challenges they address.

Fed-MNMT (Liu et al., 2023f) applies adapter-based fine-tuning for multilingual neural machine translation, significantly reducing communication overhead. It further explores parameter clustering strategies to mitigate conflicts during aggregation. **AdaFedSelecKD** (Feng et al., 2024a) performs adapter-based summarization to minimize transmitted parameters and introduces selective knowledge distillation for efficient domain-adaptive learning. **FedDAT** (Chen et al., 2024a) proposes a dual-adapter teacher framework to regularize local updates under data heterogeneity, and employs mutual knowledge distillation for effective cross-client knowledge transfer.

4.4 Selective-based Tuning

4.4.1 Preliminary

Selective-based tuning has emerged as an efficient strategy for fine-tuning LLMs by updating specific parameters of the model while keeping the majority of pre-trained weights frozen (Kornblith et al., 2019). This approach significantly reduces computational and memory overhead compared to full-parameter fine-tuning, while maintaining the model's generalization ability. Among selective-based tuning techniques, two widely adopted strategies are bias tuning (Zaken et al., 2021) and partial tuning (Houlsby et al., 2019), both of which optimize only a subset of parameters rather than the entire model.

Bias tuning updates only the bias terms of the model while keeping all other parameters frozen. Despite its simplicity, this approach has demonstrated strong performance across a variety of tasks with minimal overhead. Partial tuning generalizes this idea by allowing updates to a carefully selected subset of model parameters, such as layer normalization parameters, feed-forward network biases, or specific attention blocks. By focusing updates on the most relevant parameters, selective-based tuning methods improve training efficiency, mitigate catastrophic forgetting, and enable rapid adaptation using limited data and resources.

4.4.2 Selective Fine-tuning Methods

DP-BiTFiT (Bu et al., 2022) applies differentially private bias-term tuning in centralized training scenarios to ensure privacy-preserving adaptation while reducing resource demands. **FedPEFT** (Sun et al., 2022) shares only a small subset of model weights, such as bias parameters, significantly reducing the communication overhead. **RaFFM** (Yu et al., 2023d) introduces a resource-aware model compression framework tailored for FL, which includes salient parameter prioritization and subnetwork extraction to support dynamic model scaling across heterogeneous edge devices. Sun et al. (2024) propose a selective layer-wise fine-tuning approach to reduce training cost while preserving model performance.

4.5 Other Tuning Methods

In addition to mainstream fine-tuning strategies, several alternative approaches have been explored to optimize LLMs in FL. These methods primarily include **zeroth-order optimization** (Chen et al., 2017), **split learning** (Thapa et al., 2022), **model compression** (Choudhary et al., 2020), and **data selection** (Shen, 2024) (as shown in Figure 9). For example, **FedKSeed** (Qin et al., 2023b) employs zeroth-order optimization with a finite set of random seeds, enabling LLM fine-tuning without storing intermediate activations and reducing communication overhead. **FedBERT** (Tian et al., 2022b) combines FL with split learning to pre-train BERT in a distributed, privacy-preserving manner, achieving efficient model training across decentralized clients. **FedBiOT** (Wu et al., 2024b) compresses the LLM at the server side while allowing clients to fine-tune lightweight adapters, significantly reducing resource consumption. **FedHDS** (Qin et al., 2024) introduces a hierarchical data selection framework that identifies representative coresets for instruction tuning, minimizing redundancy at both intra- and inter-client levels to improve training efficiency in FL.

5 Datasets and Benchmarks

A comprehensive evaluation framework is essential for systematically assessing the effectiveness and generalization ability of federated fine-tuning methods. In this section, we begin by presenting widely-used fine-tuning datasets spanning multiple domains. We then detail a suite of domain-specific evaluation bench-

Table 9: A summary of representative **instruction fine-tuning datasets**. The construction methods are categorized into three types: human construct, which refers to datasets manually written or annotated by humans; model construct, which refers to data generated by prompting LLMs; and synthetic, which refers to data produced through rule-based or programmatic generation. The datasets span key domains such as general language understanding, finance, medicine, code, math, and law.

Dataset	Language	Construction Method	Domain	Description
Alpaca	English	Model Construct	General	Generated by Text-Davinci-003 with Alpaca-style instruction prompts.
Alpaca-GPT4	English	Model Construct	General	Generated by GPT-4 based on Alpaca prompts with more nuanced multi-turn instructions.
Self-Instruct	English	Human + Model Construct	General	Dataset from seed instructions using GPT-3 for improving model generalization.
UltraChat 200k	English	Model Construct	General	High-quality multi-turn dialogue dataset filtered from UltraChat.
OpenOrca	English	Model Construct	General	Dataset of $\sim$ 4.2M GPT-3.5/4 augmented FLAN examples for instruction following.
ShareGPT90K	English	Model Construct	General	Multi-turn dialogue dataset of 90K samples derived from ShareGPT for instruction following.
WizardLM Evol-Instruct V2 196k	English	Model Construct	General	Dataset of 196K samples generated via Evol-Instruct method.
Databricks Dolly 15K	English	Human Construct	General	Dataset of 15K human-generated prompt-response pairs across diverse tasks.
Baize	English	Model Construct	General	Instruction-following dialogues generated by prompting ChatGPT with user-centric queries.
OpenChat	English	Model Construct	General	Instruction-following dialogues generated using open LLMs for multi-turn alignment.
Flan-v2	English	Model Construct	General	Dataset combining Flan, P3, Super-Natural Instructions, Chain-of-thought, and Dialog tasks.
BELLE-train-0.5M-CN	Chinese	Human + Model Construct	General	Dataset of 519K Chinese samples for instruction following.
BELLE-train-1M-CN	Chinese	Human + Model Construct	General	Dataset of 917K Chinese samples for instruction following.
BELLE-train-2M-CN	Chinese	Human + Model Construct	General	Dataset of 2M Chinese samples for instruction following.
Firefly-train-1.1M	Chinese	Human Construct	General	Dataset of 1.65M Chinese samples across 23 tasks with human-written instruction templates.
Wizard-LM-Chinese-instruct-evol	Chinese	Human + Model Construct	General	Dataset of 70K samples translated from WizardLM to Chinese with GPT-generated responses.
HC3-Chinese	Chinese	Human + Model Construct	General	Chinese Human-ChatGPT QA pairs across multiple domains.
HC3	English / Chinese	Human + Model Construct	General	Human-ChatGPT QA pairs across multiple domains.
ShareGPT-Chinese-English-90k	English / Chinese	Model Construct	General	Bilingual human-machine QA dataset of 90K real user queries in Chinese and English.
FinGPT	English	Human + Model Construct	Finance	Instruction tuning data for financial tasks.
Finance-Instruct-500k	English	Human + Model Construct	Finance	Large-scale instruction tuning dataset for financial tasks.
Finance-Alpaca	English	Human + Model Construct	Finance	Instruction tuning dataset combining Alpaca and FiQA.
Financial PhraseBank	English	Human Construct	Finance	Manually annotated financial news sentences for sentiment classification.
Yahoo-Finance-Data	English	Human Construct	Finance	Financial dataset including estimates, statements, and historical prices from Yahoo Finance.
Financial-QA-10K	English	Model Construct	Finance	Contextual QA dataset for financial question answering and retrieval tasks.
Financial-Classification	English	Human Construct	Finance	Dataset combining Financial PhraseBank and Kaggle financial texts.
Twitter-Financial-News-Topic	English	Human Construct	Finance	Annotated tweets for multi-class financial topic classification tasks.
Financial-News-Articles	English	Human Construct	Finance	Financial news articles for text classification and sentiment analysis tasks.
FiQA	English	Human Construct	Finance	QA dataset from financial texts and forums covering finance topics.
Earnings-Call	English	Human Construct	Finance	QA pairs from CEO/CFO earnings calls for financial reasoning.
Doc2EDAG	Chinese	Human Construct	Finance	Financial reports annotated for event graph extraction and structuring.
Synthetic-PII-Finance-Multilingual	Multilingual	Synthetic	Finance	Financial documents with labeled PII for NER and privacy-preserving model training.
ChatDoctor-200K	English	Human + Model Construct	Medical	Instruction tuning dataset for medical QA and dialogue generation.
ChatDoctor-HealthCareMagic-100k	English	Human Construct	Medical	Real-world doctor-patient conversations.
Medical Meadow CORD-19	English	Human Construct	Medical	Instruction tuning dataset for medical literature summarization based on CORD-19.
Medical Meadow MedQA	English	Human Construct	Medical	Multiple-choice medical QA dataset derived from professional medical board exams.
HealthCareMagic-100k-en	English	Human Construct	Medical	Real doctor-patient consultations for instruction tuning.
ChatMed-Consult-Dataset	Chinese	Human + Model Construct	Medical	Instruction tuning dataset of Chinese medical consultations for QA and dialogue generation.
CMtMedQA	Chinese	Human Construct	Medical	Multi-turn medical QA dataset with real doctor-patient conversations.
DISC-Med-SFT	Chinese	Human + Model Construct	Medical	Instruction tuning dataset combining real-world dialogues and knowledge graph QA pairs.
Huatuo-26M	Chinese	Human Construct	Medical	Structured QA pairs extracted from medical encyclopedias and articles.
Huatuo26M-Lite	Chinese	Human + Model Construct	Medical	Refined subset of Huatuo-26M with ChatGPT-rewritten answers.
ShenNong-TCM-Dataset	Chinese	Human + Model Construct	Medical	QA pairs generated via entity-centric self-instruct from a TCM knowledge graph.
HuatuoGPT-sft-data-v1	Chinese	Human + Model Construct	Medical	Instruction tuning dataset combining ChatGPT-generated and real doctor-patient dialogues.
MedDialog	English / Chinese	Human Construct	Medical	Large-scale doctor-patient dialogues for medical QA and dialogue generation.
CodeAlpaca	English	Model Construct	Code	GPT-generated code instruction-following dataset based on the Alpaca style.
Code Instructions 120k Alpaca	English	Human + Model Construct	Code	Instruction tuning dataset for code generation with prompts in alpaca style.
CodeContests	English	Human Construct	Code	Dataset from competitive programming platforms, designed for program synthesis and reasoning.
CommitPackFT	English	Human Construct	Code	Filtered dataset of GitHub commits with high-quality messages for code generation tasks.
ToolBench	English	Human + Model Construct	Code	Instruction tuning dataset for multi-tool API usage scenarios.
CodeParrot-Clean	English	Human Construct	Code	Deduplicated and filtered dataset of Python files from GitHub for code generation.
The Stack v2 Dedup	English	Human Construct	Code	Large-scale deduplicated dataset of source code from 600+ programming languages.
CodeSearchNet	English	Human Construct	Code	Code-language pairs for retrieval and semantic search across six programming languages.
CodeForces-CoTs	English	Human + Model Construct	Code	A dataset consists of 10k CodeForces problems with reasoning traces generated by DeepSeek R1.
CodeXGLUE Code Refinement	English	Human Construct	Code	Dataset of buggy and fixed Java functions for code refinement.
GSM8K	English	Human Construct	Math	A dataset of 8.5k grade school-level arithmetic word problems with detailed step-by-step solutions.
CoT-GSM8k	English	Human + Model Construct	Math	Extended version of GSM8K with chain-of-thought reasoning steps.
MathInstruct	English	Human + Model Construct	Math	Dataset combining CoT and program-of-thought rationales across diverse mathematical fields.
MetaMathQA	English	Human + Model Construct	Math	Multi-perspective question augmentations from GSM8K and MATH.
OpenR1-Math-220k	English	Human + Model Construct	Math	Dataset of 220k math problems with multiple reasoning traces generated by DeepSeek R1.
Hendrycks MATH Benchmark	English	Human Construct	Math	Dataset of 12.5K high school competition math problems with step-by-step solutions.
DeepMind Mathematics Dataset	English	Synthetic	Math	Dataset of algorithmically generated math problems across various topics.
OpenMathInstruct-1	English	Human + Model Construct	Math	Dataset of 1.8M math problems with code-interpreter solutions generated by Mixtral-8x7B.
Orca-Math Word Problems 200k	English	Synthetic	Math	Synthetic dataset of 200K grade school math word problems with GPT-4 Turbo solutions.
DAPO-Math-17k	English	Human + Model Construct	Math	Dataset of 17K diverse math problems for mathematical reasoning.
Big-Math-RL-Verified	English	Human + Model Construct	Math	Dataset of 251K math problems with verifiable answers, curated for reinforcement learning.
BELLE-math-zh	Chinese	Human + Model Construct	Math	Dataset of Chinese elementary school math problems with step-by-step solutions.
MathInstruct-Chinese	Chinese	Model Construct	Math	Chinese version of MathInstruct; contains instruction-style math problems in Chinese.
Legal-QA-v1	English	Human Construct	Law	Dataset of 3.7K legal question-answer pairs sourced from legal forums.
Pile of Law	English	Human Construct	Law	Dataset for legal-domain tasks.
CUAD	English	Human Construct	Law	Expert-annotated dataset of 26K legal contract question-answer pairs across 41 clause categories.
LEDGAR	English	Human Construct	Law	Dataset of 1.45M contract clauses labeled by legal experts.
DISC-Law-SFT	Chinese	Human + Model Construct	Law	Comprehensive instruction tuning dataset covering diverse legal tasks.
Law-GPT-zh	Chinese	Human Construct	Law	Dataset of legal sentence pairs for training sentence embedding models in legal domain.
Lawyer LLaMA Data	Chinese	Human + Model Construct	Law	Dataset for Chinese legal tasks including legal consultation and bar exam question answering.

marks that enable consistent, fine-grained, and standardized assessment of FedLLM performance across diverse and heterogeneous task settings.

5.1 Instruction Fine-tuning Datasets

Table 9 presents a comprehensive collection of instruction fine-tuning datasets spanning several key domains, including general language understanding, finance, medicine, code, math, and law. For each domain, we select representative datasets from Hugging Face that are widely adopted by the community and closely aligned

with real-world FedLLM applications. Building on this overview, we next provide a detailed introduction to the datasets within each domain.

5.1.1 General Instruction Fine-tuning Datasets

General instruction fine-tuning datasets are primarily designed to improve the overall instruction-following capability of LLMs across a wide range of tasks and domains (Zhou et al., 2023). These datasets serve as a foundation for aligning LLMs with human intent, and are particularly useful in federated settings where clients may engage in diverse yet general-purpose interactions.

For English instruction fine-tuning, representative datasets include Alpaca (Taori et al., 2023), which is generated by Text-Davinci-003 with Alpaca-style instruction prompts, and Alpaca-GPT4 (Peng et al., 2023), which builds on this with GPT-4-generated, multi-turn dialogues to improve linguistic nuance and contextual coherence. Other commonly used datasets include Self-Instruct (Wang et al., 2022b), UltraChat 200k (Ding et al., 2023a), OpenOrca (Lian et al., 2023), ShareGPT-90K[1], WizardLM Evol-Instruct V2 196k (Xu et al., 2023a), Databricks Dolly 15K (Conover et al., 2023), Baize (Xu et al., 2023b), OpenChat (Wang et al., 2023a), and Flan-v2[2]. A detailed comparison of these datasets is provided in Table 9.

For Chinese instruction tuning, representative datasets include BELLE-train-0.5M-CN (Ji et al., 2023a), BELLE-train-1M-CN (Ji et al., 2023a), and BELLE-train-2M-CN (Ji et al., 2023a), which provide progressively scaled corpora designed to improve model alignment with Chinese user instructions. Additional resources such as Firefly-train-1.1M (Yang, 2023), Wizard-LM-Chinese-Instruct-Evol (Xu et al., 2023a), and HC3-Chinese (Guo et al., 2023a) support a wide range of instruction tasks, including reasoning, question answering, and domain-specific knowledge modeling.

Bilingual instruction datasets are also included to support multilingual and cross-lingual instruction tuning in federated contexts. HC3 (Guo et al., 2023a) offers both human- and model-generated responses in English and Chinese for evaluating factual consistency and detection capabilities. ShareGPT-Chinese-English-90k (shareAI, 2023) provides high-quality, bilingual conversations, making it particularly suitable for instruction tuning in multilingual FedLLM scenarios.

5.1.2 Financial Domain Instruction Fine-Tuning Datasets

Instruction fine-tuning datasets in the financial domain are tailored to equip language models with specialized knowledge and reasoning capabilities relevant to financial tasks (Li et al., 2023h). In the context of FedLLM, such datasets are particularly valuable for enabling privacy-preserving and institution-specific applications, including investment recommendation, market sentiment analysis, financial reporting assistance, and regulatory compliance. These use cases often involve sensitive and proprietary data that cannot be centrally aggregated due to privacy, confidentiality, or regulatory constraints (Byrd & Polychroniadou, 2020), making federated fine-tuning an ideal solution.

For English instruction tuning, representative datasets include FinGPT (Zhang et al., 2023a), which provides a large corpus of financial question-answering pairs and document summaries tailored to real-world financial analysis. Finance-Instruct-500k (Flowers, 2025) and Finance-Alpaca[3] extend general instruction-tuning formats to financial scenarios, offering instruction–response pairs related to stock prediction, portfolio analysis, and macroeconomic commentary. Additional datasets such as Financial PhraseBank (Malo et al., 2014), Yahoo-Finance-Data[4], Financial-QA-10K[5], Financial-Classification[6], Twitter-Financial-News-Topic[7], Financial-News-Articles[8], and FiQA (Maia et al., 2018) cover a broad spectrum of financial tasks including sentiment classification, time-series event extraction, and financial question answering. Transcripts from earnings calls further support fine-tuning for multi-turn, dialogue-based financial reasoning.

[1] https://huggingface.co/datasets/liyucheng/ShareGPT90K
[2] https://huggingface.co/datasets/SirNeural/flan_v2
[3] https://huggingface.co/datasets/gbharti/finance-alpaca
[4] https://huggingface.co/datasets/bwzheng2010/yahoo-finance-data
[5] https://huggingface.co/datasets/virattt/financial-qa-10K
[6] https://huggingface.co/datasets/nickmuchi/financial-classification
[7] https://huggingface.co/datasets/zeroshot/twitter-financial-news-topic
[8] https://huggingface.co/datasets/ashraq/financial-news-articles

For Chinese financial applications, Doc2EDAG (Zheng et al., 2019) provides a rich dataset for event detection and argument generation from financial documents, supporting instruction-style tasks. In addition, the Synthetic-PII-Finance-Multilingual dataset (Watson et al., 2024) includes multi-language synthetic financial records annotated with privacy-related attributes. A comparative summary of these finance-related datasets is presented in Table 9.

5.1.3 Medical Domain Instruction Fine-Tuning Datasets

Instruction fine-tuning datasets in the medical domain are designed to enable LLMs to perform tasks such as medical reasoning, patient interaction, and clinical decision support (Zhang et al., 2023h). Within the context of FedLLM, these datasets are particularly relevant for privacy-preserving healthcare applications, where sensitive patient data is inherently distributed across hospitals, clinics, and personal health devices, and cannot be centrally aggregated due to strict privacy regulations (Nguyen et al., 2022). Federated fine-tuning with medical instruction data empowers models to generalize across diverse clinical intents while preserving the confidentiality and heterogeneity of local medical records, thereby supporting robust and compliant deployment in real-world healthcare settings.

For English-language datasets, ChatDoctor-200K (Li et al., 2023i) and ChatDoctor-HealthCareMagic-100k (Li et al., 2023i) contain medical dialogues derived from professional consultation platforms, facilitating multi-turn reasoning and symptom analysis. The Medical Meadow (Han et al., 2023) suite offers curated datasets from sources such as CORD-19[9] and MedQA[10], targeting tasks like biomedical question answering, literature summarization, and clinical fact verification. Additionally, HealthCareMagic-100k-en[11] supports English patient–doctor interactions across various medical specialties.

For Chinese medical instruction tuning, a rich set of datasets has been developed to address the unique linguistic and clinical characteristics of Chinese healthcare scenarios. Notable examples include ChatMed-Consult-Dataset (Zhu et al., 2023b) and CMtMedQA (Yang et al., 2024a), which focus on consultation-style QA pairs. DISC-Med-SFT (Bao et al., 2023), Huatuo-26M (Li et al., 2023d), Huatuo26M-Lite[12], and HuatuoGPT-sft-data-v1[13] offer large-scale instruction–response pairs in clinical medicine, public health, and disease treatment. Traditional Chinese Medicine (TCM) is also covered through datasets like ShenNong-TCM-Dataset (Zhu & Wang, 2023), enabling LLMs to support specific diagnostic and treatment tasks.

The MedDialog (Zeng et al., 2020) dataset provides bilingual (English and Chinese) medical dialogues, supporting cross-lingual instruction tuning and evaluation in federated medical environments where linguistic diversity and clinical protocol variance are prevalent. A comparative summary of these medical-specific datasets, including language, construction method, and description, is provided in Table 9.

5.1.4 Code Domain Instruction Fine-Tuning Datasets

Instruction fine-tuning datasets in the code domain are curated to improve a language model's ability to understand, generate, and reason about source code across various programming languages and tasks (Muennighoff et al., 2023). In the context of FedLLM, such datasets are particularly valuable for enabling privacy-preserving applications like on-device programming assistants, secure code generation within enterprise environments, and personalized developer support. Given that source code repositories often contain proprietary algorithms, sensitive business logic, or embedded credentials, federated fine-tuning offers a compelling alternative to centralized training on raw code, allowing organizations to harness LLM capabilities without compromising code confidentiality.

Representative datasets include CodeAlpaca (Chaudhary, 2023) and Code Instructions 120k Alpaca[14], which extend the Alpaca instruction format to software engineering tasks such as debugging, function generation, and refactoring. CodeContests (Li et al., 2022c) focuses on competitive programming tasks and provides

[9] https://huggingface.co/datasets/medalpaca/medical_meadow_cord19
[10] https://huggingface.co/datasets/medalpaca/medical_meadow_medqa
[11] https://huggingface.co/datasets/wangrongsheng/HealthCareMagic-100k-en
[12] https://huggingface.co/datasets/FreedomIntelligence/Huatuo26M-Lite
[13] https://huggingface.co/datasets/FreedomIntelligence/HuatuoGPT-sft-data-v1
[14] https://huggingface.co/datasets/iamtarun/code_instructions_120k_alpaca

instruction–response pairs related to algorithmic problem solving. CommitPackFT (Muennighoff et al., 2023) offers commit-message generation tasks based on source code diffs, reflecting real-world software maintenance scenarios. ToolBench (Qin et al., 2023a) is designed to help models learn tool-augmented code generation through instruction-following examples.

Several large-scale pretraining and fine-tuning datasets are also widely used for code instruction alignment. CodeParrot-Clean[15] and The Stack v2 Dedup (Kocetkov et al., 2022) provide diverse and deduplicated code corpora across multiple programming languages, while CodeSearchNet (Husain et al., 2019) and CodeXGLUE (Lu et al., 2021) support retrieval, summarization, and translation tasks with instruction-style prompts. CodeForces-CoTs (Penedo et al., 2025) incorporates chain-of-thought annotations for programming tasks, supporting more explainable and step-wise code generation. These code-specific datasets play a vital role in developing FedLLMs that can support privacy-sensitive, domain-specific programming environments. A comparative summary of these code-related datasets is presented in Table 9.

5.1.5 Math Domain Instruction Fine-Tuning Datasets

Instruction fine-tuning datasets in the math domain are developed to enhance a language model's proficiency in mathematical reasoning, symbolic computation, and step-by-step problem solving (Tang et al., 2024b). These datasets are particularly valuable in FedLLM scenarios such as personalized math tutoring, intelligent educational platforms, and localized STEM applications, where sensitive student information or institution-specific curricular content must remain on-device. Fine-tuning LLMs in the math domain poses unique challenges, as it requires not only a strong grasp of language but also precise logical inference and numerical accuracy—skills essential for generating correct and interpretable mathematical solutions.

For English-language datasets, GSM8K (Cobbe et al., 2021b) is a widely used benchmark for grade-school math word problems with detailed rationales. CoT-GSM8K[16] augments it with chain-of-thought explanations to support intermediate reasoning steps. Datasets such as MathInstruct (Yue et al., 2023c), MetaMathQA (Yu et al., 2023b), OpenR1-Math-220k (Allal et al., 2025), and Orca-Math Word Problems 200k (Mitra et al., 2024) provide high-quality, instruction-based math problems covering arithmetic, algebra, and word problem solving. The Hendrycks MATH Benchmark (Hendrycks et al., 2021c) and DeepMind Mathematics[17] Dataset offer more advanced, competition-style problems suitable for evaluating formal mathematical reasoning. Recent datasets like DAPO-Math-17k (Yu et al., 2025), Big-Math-RL-Verified (Albalak et al., 2025), and OpenMathInstruct-1 (Toshniwal et al., 2024) integrate reinforcement signals, verifiable proofs, or multi-step derivations, further pushing the boundaries of instruction-aligned mathematical LLMs.

For Chinese-language instruction tuning datasets, BELLE-math-zh[18] and MathInstruct-Chinese[19] provide diverse mathematical problems adapted to Chinese curricula and linguistic structures. These datasets facilitate the training and evaluation of FedLLMs in multilingual educational contexts, supporting privacy-preserving and culturally contextualized math assistance. A comparative summary of these math-focused instruction tuning datasets is provided in Table 9.

5.1.6 Legal Domain Instruction Fine-Tuning Datasets

Instruction fine-tuning datasets in the legal domain are designed to improve a language model's capability in legal reasoning, contract analysis, statute interpretation, and other tasks that require deep, domain-specific understanding of legal language and structure (Yue et al., 2023a). In the context of FedLLM, these datasets are particularly important for enabling decentralized legal assistance systems, confidential contract review, and on-device compliance monitoring—applications where legal data is often sensitive, jurisdiction-bound, and subject to strict confidentiality constraints. As centralized training on legal documents is frequently infeasible due to regulatory and privacy concerns, federated fine-tuning offers a promising approach to leveraging LLMs in legal settings without compromising data security or legal integrity.

[15] https://huggingface.co/datasets/codeparrot/codeparrot-clean
[16] https://huggingface.co/datasets/Dahoas/cot_gsm8k
[17] https://huggingface.co/datasets/di-zhang-fdu/DeepMind_Mathematics_QA
[18] https://huggingface.co/datasets/frankminors123/belle-math-zh
[19] https://huggingface.co/datasets/ALmonster/MathInstruct-Chinese

For English-language instruction tuning, Legal-QA-v1[20] provides question–answer pairs covering legal concepts and procedures across various subfields of law. Pile of Law (Gao et al., 2020) is a large-scale corpus of U.S. legal documents—including court opinions, contracts, and regulations—that supports open-ended legal instruction tuning. CUAD (Hendrycks et al., 2021b) focuses on contract understanding, with annotated question–answer pairs tailored for clause extraction and risk analysis. LEDGAR[21] contains a large set of contractual clauses categorized into fine-grained legal functions, useful for classification and retrieval tasks in instruction-based formats.

For Chinese-language legal modeling, DISC-Law-SFT (Yue et al., 2023b) offers supervised instruction–response pairs across a wide spectrum of Chinese legal domains, including civil law, criminal law, and administrative law. Law-GPT-zh[22] consolidates multiple sources of Chinese legal texts into an instruction tuning format to support legal consultation and statutory reasoning. Lawyer LLaMA Data (Huang et al., 2023a) further augments legal dialogue capabilities with simulated lawyer–client conversations, making it well-suited for on-device legal assistants in federated environments. A comparative overview of these law-specific instruction tuning datasets is provided in Table 9.

5.2 Evaluation Benchmarks

5.2.1 General Evaluation Benchmarks

General-purpose evaluation benchmarks play a critical role in systematically assessing the instruction-following ability, reasoning competence, and overall robustness of LLMs across a wide range of tasks and domains (Lou et al., 2024). In the context of FedLLM, such benchmarks are especially valuable for evaluating model generalization under heterogeneous data distributions, identifying robustness gaps in decentralized training settings, and facilitating consistent comparisons between personalized and globally aggregated models. Existing benchmarks can be broadly categorized into four groups: (i) general reasoning and instruction-following, (ii) robustness, alignment, and meta-evaluation, (iii) multilingual and Chinese-specific benchmarks, and (iv) long-context understanding. These benchmarks provide a foundation for rigorous and reproducible evaluation of FedLLM across heterogeneous and dynamic environments.

(i) General reasoning and instruction-following. This category includes evaluation benchmarks such as MMLU (Hendrycks et al., 2020), BIG-bench (Srivastava et al., 2022), DROP (Dua et al., 2019), CRASS (Frohberg & Binder, 2021), and ARC (Clark et al., 2018), which assess multitask knowledge, discrete reasoning, and science question answering. AGIEval (Zhong et al., 2023), M3Exam (Zhang et al., 2023f), and SCIBENCH (Wang et al., 2023d) extend evaluation to standardized exams and college-level math, physics, and chemistry. Instruction-following quality is addressed by Vicuna Evaluation[23], MT-Bench (Zheng et al., 2023a), AlpacaEval (Dubois et al., 2023), Chatbot Arena (Zheng et al., 2023a), and PandaLM (Wang et al., 2023f). Datasets like HellaSwag (Zellers et al., 2019) and TruthfulQA (Lin et al., 2021) focus on commonsense inference and factual accuracy, respectively.

Several benchmarks target more specialized reasoning abilities: ScienceQA (Lu et al., 2022a) evaluates multimodal scientific question answering; Chain-of-Thought Hub (Fu et al., 2023) focuses on step-wise reasoning; NeuLR (Xu et al., 2023c) benchmarks deductive, inductive, and abductive reasoning; ALCUNA (Yin et al., 2023) assesses generalization to novel knowledge; LMExamQA (Bai et al., 2023c) tests models on recall, understanding, and analysis across over academic questions. Benchmarks like SocKET (Choi et al., 2023) and Choice-75 (Hou et al., 2023) address social knowledge and decision-making in scripted scenarios, respectively. Broad-scoped platforms such as HELM (Liang et al., 2022) and OpenLLM[24] integrate multiple datasets and offer normalized aggregate scores across diverse metrics and tasks.

(ii) Robustness, alignment, and meta-evaluation. To simulate the variability and noise of federated environments, benchmarks such as BOSS (Yuan et al., 2023), GLUE-X (Yang et al., 2022), PromptBench (Zhu et al., 2023a), and DynaBench (Kiela et al., 2021) test model robustness to input distribution shifts, prompt

[20] https://huggingface.co/datasets/dzunggg/legal-qa-v1
[21] https://huggingface.co/datasets/coastalchp/ledgar
[22] https://huggingface.co/datasets/sentence-transformers/law-gpt
[23] https://github.com/lm-sys/vicuna-blog-eval
[24] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Table 10: Overview of general-domain evaluation benchmarks. Benchmarks specifically designed for *long-context* evaluation are highlighted. For each benchmark, we summarize its primary evaluation objective and the main evaluation criteria.

Benchmark	Domain	Evaluation Objective	Main Evaluation Criteria
MMLU	General	Evaluate multitask language understanding across 57 subjects	Multitask accuracy
BIG-bench	General	Evaluate advanced reasoning capabilities	Model performance and calibration
DROP	General	Evaluate discrete reasoning over paragraphs	Exact match and F1 score
CRASS	General	Evaluate counterfactual reasoning ability in LLMs	Multiple-choice accuracy
ARC	General	Assess science reasoning at grade-school level	Multiple-choice accuracy
AGIEval	General	Evaluate foundation models on human-centric standardized exam tasks	Multi-task accuracy across disciplines
M3Exam	General	Evaluate multilingual, multimodal, and multilevel reasoning across real exam questions	Multiple-choice accuracy
SCIBENCH	General	Evaluate college-level scientific problem-solving in math, physics, and chemistry	Open-ended accuracy and skill-specific error attribution
Vicuna Evaluation	General	Evaluate instruction-following quality in chat settings	Human/GPT-4 preference comparison accuracy
MT-Bench	General	Evaluate multi-turn conversational and instruction-following capabilities	Winrate judged by GPT-4
AlpacaEval	General	Evaluate instruction-following performance via LLM-based auto-annotation	Length-controlled win rate correlated with human preference accuracy
Chatbot Arena	General	Evaluate LLMs via human-voted battles using Elo rating	Elo score and head-to-head win rate
PandaLM	General	Evaluate instruction-following quality and hyperparameter impact	Winrate judged by PandaLM
HellaSwag	General	Evaluate commonsense inference by selecting the most plausible continuation	Multiple-choice accuracy
TruthfulQA	General	Evaluate LLM truthfulness and avoidance of imitative falsehoods	Truthfulness rate and MC accuracy
ScienceQA	General	Evaluate multimodal scientific reasoning and explanation generation	Multiple-choice accuracy and explanation quality
Chain-of-Thought Hub	General	Evaluate LLMs' multi-step reasoning across diverse domains using CoT prompting	Few-shot CoT accuracy
NeuLR	General	Evaluate deductive, inductive, and abductive reasoning capabilities of LLMs	Multi-dimensional evaluation (e.g., accuracy, self-awareness)
ALCUNA	General	Evaluate LLMs' ability to comprehend and reason over novel knowledge	Accuracy on 84,351 queries
LMExamQA	General	Evaluate LLMs' performance across knowledge recall, understanding, and analysis	Accuracy on 10,090 questions
SocKET	General	Evaluate LLMs' sociability and understanding of social knowledge	Accuracy and other metrics
Choice-75	General	Evaluate LLMs' decision reasoning ability within scripted scenarios	Accuracy on binary multiple-choice questions
HELM	General	Evaluate LMs via multi-metric scenarios	Composite performance ranking (normalized scores across metrics)
OpenLLM	General	Evaluate open-style reasoning across multiple benchmarks	Normalized accuracy aggregate across tasks
BOSS	General	Evaluate out-of-distribution robustness across NLP tasks	OOD accuracy drop and ID–OOD performance correlation
GLUE-X	General	Evaluate out-of-distribution robustness	Average OOD accuracy drop relative to ID performance
PromptBench	General	Evaluate robustness, prompt-engineering and dynamic evaluation capabilities	Adversarial success rate, robustness, prompt-variant performance
DynaBench	General	Evaluate robustness via dynamic, human-in-loop adversarial data collection	Model error rate on human-crafted challenge examples
KoLA	General	Evaluate world knowledge across 19 evolving tasks	self-contrast calibration
CELLO	General	Evaluate following complex, real-world instructions with multiple constraints	Multi-criteria compliance rate judged by automated metrics
LLMEval	General	Evaluate LLM evaluators' ability to assess instruction-following quality	Meta-evaluator accuracy
Xiezhi	General	Evaluate holistic domain knowledge across 516 disciplines	Mean Reciprocal Rank (MRR)
C-Eval	General	Evaluate Chinese domain knowledge and reasoning across 52 disciplines	Multiple-choice accuracy
BELLE-eval	General	Evaluate Chinese instruction-following and multi-skill capabilities	GPT-4 adjudicated win-rate and per-task quality scores
SuperCLUE	General	Evaluate Chinese instruction-following with human preference alignment	GPT-4 adjudicated win-rate and Elo scores
M3KE	General	Evaluate Chinese LLMs' knowledge across 71 disciplines and 4 education levels	Multitask accuracy in zero- and few-shot settings
BayLing-80	General	Evaluate cross-lingual and conversational capabilities	GPT-4 adjudicated win-rate
MMCU	General	Evaluate multitask Chinese understanding across different domains	Multitask accuracy (zero-/few-shot)
C-CLUE	General	Evaluate classical Chinese NER and relation extraction capabilities	Weighted F1 score for NER and RE tasks
LongBench	General (long-context)	Evaluate bilingual long-context understanding across multiple real-world tasks	Accuracy and generation quality
L-Eval	General (long-context)	Evaluate long-context reasoning and task performance over documents up to 60K tokens	Multi-metric assessment
InfinityBench	General (long-context)	Evaluate long-context understanding across tasks up to 2M tokens in length	Accuracy and task-specific metrics
Marathon	General (long-context)	Evaluate long-context understanding across multi-domain and multi-task settings	Accuracy, F1, ROUGE-L, EM across tasks and context lengths
LongEval	General (long-context)	Evaluate LLMs' effectiveness in long-context retrieval tasks	Accuracy in broad-topic and fine-grained passage retrieval
BABILong	General (long-context)	Evaluate long-context reasoning in haystack-like settings	Accuracy on long-context reasoning tasks
DetectiveQA	General (long-context)	Evaluate long-context reasoning through detective novel comprehension	Instruction-following accuracy and winrate judged by PandaLM
NoCha	General (long-context)	Evaluate narrative comprehension and long-range coreference resolution in literary texts	Exact match and accuracy
Loong	General (long-context)	Evaluate LLMs' ability to perform multi-document reasoning and extended-context question answering	Answer accuracy, context utilization rate, and document coverage score
TCELongBench	General (long-context)	Evaluate temporal reasoning over complex, long event narratives	Temporal event ordering accuracy and event query answering accuracy
DENIAHL	General (long-context)	Evaluate the influence of in-context features on LLMs' needle-in-a-haystack abilities	Accuracy on long-context retrieval and reasoning tasks
LongMemEval	General (long-context)	Benchmark LLMs' ability to retain and utilize long-term interactive memory across dialogue sessions	Memory retention accuracy, retrieval success rate, and context utilization score
Long2RAG	General (long-context)	Evaluate long-form generation and retrieval grounding under extended input lengths	Key point recall, LLM-based coherence, factuality, and answer relevance
L-CiteEval	General (long-context)	Evaluate LLMs' ability to utilize citations and contextual evidence	Citation recall, LLM-based factuality and faithfulness scoring
LIFBENCH	General (long-context)	Evaluate instruction-following performance and robustness in long-context scenarios	Instruction-following accuracy and response stability
LongReason	General (long-context)	Evaluate long-context reasoning via synthetic context expansion	Instruction-following accuracy and reasoning correctness
BAMBOO	General (long-context)	Evaluate long-text modeling capacity across diverse real-world tasks	Instruction-following accuracy and winrate judged by PandaLM
ETHIC	General (long-context)	Evaluate instruction-following on long-context tasks with high information coverage	Exact match, F1, and consistency/completeness
LooGLE	General (long-context)	Evaluate LLMs' ability to understand and reason over long contexts	Accuracy on inputs averaging 20K words
HELMET	General (long-context)	Evaluate instruction-following, retrieval, reasoning, summarization, and long-context understanding	Task-specific automatic metrics and human evaluation
HoloBench	General (long-context)	Evaluate LLMs' holistic reasoning ability over database-style textual inputs	Execution accuracy and reasoning consistency on DB-style tasks
LOFT	General (long-context)	Evaluate whether long-context LLMs can replace retrieval-augmented methods in complex tasks	EM/F1, SQL exec. accuracy, LLM-based factuality and reasoning coherence
Lv-Eval	General (long-context)	Evaluate long-context comprehension across five input length levels (up to 256k tokens)	Exact match, ROUGE-L, LLM-based factuality and relevance scoring
ManyICLBench	General (long-context)	Evaluate many-shot in-context learning capabilities under extended context lengths	Average accuracy across tasks under different context lengths
ZeroSCROLLS	General (long-context)	Evaluate zero-shot inference capabilities of LLMs on diverse long-text tasks	Accuracy on inputs averaging 10K words
LongICLBench	General (long-context)	Evaluate long-context LLMs' capability in in-context learning under extended input lengths	Winrate judged by PandaLM
LIBRA	General (long-context)	Evaluate long-context understanding and instruction-following in Russian	Accuracy, BLEU, LLM-based faithfulness and coherence scoring

perturbations, and adversarial examples. KoLA (Yu et al., 2023a) emphasizes world knowledge calibration, while CELLO (He et al., 2024) further tests compliance under real-world constraints. LLMEval (Zhang et al., 2023g) functions as a meta-evaluation benchmark to assess the consistency and reliability of LLM-based evaluators, an important consideration when deploying automated evaluation in federated settings.

(iii) Multilingual and Chinese-specific benchmarks. Benchmarks such as Xiezhi (Gu et al., 2024b), C-Eval (Huang et al., 2023b), BELLE-eval (Ji et al., 2023b), SuperCLUE (Xu et al., 2023d), M3KE (Liu et al., 2023a), BayLing-80 (Zhang et al., 2023d), MMCU (Zeng, 2023), and C-CLUE[25] provide a diverse evaluation landscape for Chinese and multilingual LLMs. These benchmarks span topics ranging from academic disciplines to sociocultural reasoning, using metrics including GPT-4 preference scoring, multitask accuracy, F1 scores, and normalized Elo rankings.

(iv) Long-context understanding. Long-context reasoning is critical for FedLLM applications involving document-intensive tasks, extended dialogue, and memory retention. Benchmarks such as LongBench (Bai et al., 2023b), L-Eval (An et al., 2023), InfinityBench (Zhang et al., 2024h), Marathon (Zhang et al., 2023b), LongEval (Li et al., 2023a), and BABILong (Kuratov et al., 2024) form the backbone of this category, measuring comprehension, retrieval, and reasoning over contexts up to 2 million tokens. Further specialized

[25]https://github.com/jizijing/C-CLUE

benchmarks include: 1) Narrative and temporal reasoning: DetectiveQA (Xu et al., 2024d), NoCha (Karpinska et al., 2024), Loong (Wang et al., 2024b) and TCELongBench (Zhang et al., 2024k) focus on long-range narrative understanding and event-based temporal reasoning in complex textual sequences. 2) Retrieval and memory evaluation: DENIAHL (Dai et al., 2024), LongMemEval (Wu et al., 2024a), Long2RAG (Qi et al., 2024b), and L-CiteEval (Tang et al., 2024a) assess in-context feature sensitivity, long-term memory utilization, retrieval grounding, and the model's ability to incorporate external citations.

3) Instruction-following and generation under long input: LIFBENCH (Wu et al., 2024e), LongReason (Ling et al., 2025), BAMBOO (Dong et al., 2023b), ETHIC (Lee et al., 2024b), and LooGLE (Li et al., 2023e) evaluate multi-criteria instruction-following accuracy, response stability, and coherence under extended prompts. HELMET (Yen et al., 2024) further integrates summarization, retrieval, and reasoning in unified evaluation pipelines, supporting both automatic and human metrics. 4) Database-style and structured input reasoning: HoloBench (Maekawa et al., 2024) and LOFT (Lee et al., 2024a) benchmark the ability to perform complex reasoning over structured or database-like inputs, including table querying, execution accuracy, and factual consistency, especially in scenarios where retrieval-augmented methods are substituted by long-context modeling. 5) Scaling and generalization with longer context: Lv-Eval (Yuan et al., 2024b), ManyICLBench (Zou et al., 2024), ZeroSCROLLS (Shaham et al., 2023), and LongICLBench (Li et al., 2024f) examine the scalability of in-context learning and multi-task alignment as input length increases, making them valuable tools for analyzing the performance ceiling of long-context FedLLMs. 6) Cross-lingual long-context evaluation: LIBRA (Churin et al., 2024) tests instruction-following and long-form coherence in Russian, contributing to the evaluation of multilingual long-context capabilities. These long-context benchmarks are crucial for validating the scalability and persistence of FedLLMs, especially in environments requiring private document analysis, extended user sessions, or continual knowledge tracking. Table 10 presents a comprehensive summary of the general evaluation benchmarks discussed above.

5.2.2 Financial Domain Evaluation Benchmarks

Finance-specific evaluation benchmarks are crucial for assessing the domain alignment, factual accuracy, and reasoning capabilities of LLMs in high-stakes financial contexts (Li et al., 2023h). In federated settings, where sensitive data from banks, asset managers, and regulatory bodies cannot be centralized due to confidentiality and compliance constraints, these benchmarks serve as vital tools for evaluating model performance under decentralized, privacy-preserving, and task-diverse conditions. Effective financial evaluation must encompass a broad range of tasks—including open-book question answering, multi-step quantitative reasoning, and document classification—while also accounting for linguistic nuances, regulatory requirements, and the structural complexity of financial texts. Such benchmarks are indispensable for ensuring the robustness and reliability of FedLLM in real-world financial applications.

(i) Multi-task and agent-style financial evaluation. Several benchmarks offer broad-spectrum evaluation across multiple financial tasks, aligning well with FedLLM's need to support diverse client use cases (e.g., auditing, compliance, trading). FinBen (Xie et al., 2024) evaluates holistic financial reasoning over 24 tasks using automatic metrics, retrieval-augmented generation accuracy, and expert human judgment. PIXIU (Xie et al., 2023), FLUE (Shah et al., 2022), and BBF-CFLEB (Lu et al., 2023a) benchmark LLMs on multi-task setups including sentiment classification, QA, event detection, and stock prediction. CFinBench (Nie et al., 2024) and SuperCLUEFin (Xu et al., 2024a) extend this paradigm to Chinese financial scenarios, assessing models across regulatory knowledge, certification preparation, and real-world task instructions using multi-type question formats. ICE-PIXIU (Xie et al., 2023) further supports bilingual Chinese–English evaluation, suitable for cross-regional FedLLM deployments. FLARE-ES (Zhang et al., 2024g) enables bilingual Spanish–English testing to evaluate cross-lingual transfer and domain-specific reasoning.

(ii) Task-specific capability evaluation. Fine-grained benchmarks evaluate specific financial NLP capabilities that are critical for real-world FedLLM deployment, including document question answering, information extraction, and numerical reasoning. FinanceBench (Islam et al., 2023) targets open-book question answering using real-world, company-related financial documents, with a focus on factual correctness and evidence alignment—a critical requirement for FedLLM deployed in compliance, auditing, and enterprise document analysis scenarios. FiNER-ORD (Shah et al., 2023) and FinRED (Sharma et al., 2022) evaluate financial Named Entity Recognition and relation extraction from news and earnings transcripts—key for localized

Table 11: Overview of evaluation benchmarks across five specialized domains: **finance**, **medicine**, **code**, **math**, and **law**. The table summarizes representative benchmarks along with their primary evaluation objectives, target domains, and main evaluation criteria.

Benchmark	Domain	Evaluation Objective	Main Evaluation Criteria
FinBen	Finance	Evaluate holistic financial capabilities across 24 tasks	Automatic metrics, agent/RAG performance, and human expert assessment
PIXIU	Finance	Evaluate LLMs across multiple financial NLP tasks	Sentiment accuracy, QA accuracy, stock prediction F1
FLUE	Finance	Evaluate diverse financial NLP competencies	Accuracy, F1, nDCG
BBF-CFLEB	Finance	Evaluate LLMs on Chinese financial language understanding and generation across six task types	Rouge, F1, and accuracy
CFinBench	Finance	Evaluate Chinese financial knowledge across subjects, certifications, practice, and legal compliance	Accuracy across single-choice, multiple-choice, and judgment questions
SuperCLUEFin	Finance	Evaluate Chinese financial assistant capabilities	Win-rate and multi-criteria performance
ICE-PIXIU	Finance	Evaluate bilingual (Chinese–English) financial reasoning and analysis capabilities	Task-specific accuracy and bilingual win-rate
FLARE-ES	Finance	Evaluate bilingual Spanish–English financial reasoning	Task-specific accuracy and cross-lingual transfer win-rate
FinanceBench	Finance	Evaluate financial open-book QA using real-world company-related questions	Factual correctness and evidence alignment
FiNER-ORD	Finance	Evaluate financial NER capability in financial texts	Entity F1
FinRED	Finance	Evaluate financial relation extraction performance on news and earnings transcripts	F1, Entity F1
FinQA	Finance	Evaluate multi-step numerical reasoning over financial reports with structured evidence	EM Accuracy
BizBench	Finance	Evaluate quantitative reasoning on realistic financial problems	Numeric EM, code execution pass rate, QA accuracy
EconLogicQA	Finance	Evaluate economic sequential reasoning across multi-event scenarios	Multiple-choice accuracy
FinEval	Finance	Evaluate Chinese financial domain knowledge and reasoning	Multiple-choice accuracy and task-level weighted scores
CFBenchmark	Finance	Evaluate Chinese financial assistant capabilities	Win-rate and task-specific metrics
BBT-Fin	Finance	Evaluate Chinese financial language understanding and generation	Accuracy, F1, ROUGE
Hirano	Finance	Evaluate Japanese financial language understanding	Multiple-choice accuracy and macro-F1 scores
MultiFin	Finance	Evaluate multilingual financial topic classification	F1, Multi-class accuracy
DocFinQA	Finance (long-context)	Evaluate long-context financial reasoning over documents like financial reports	EM and F1 for multi-step answer prediction, and reasoning accuracy
FinTextQA	Finance (long-context)	Evaluate long-form financial question answering with long textual context	Answer accuracy, BLEU, and ROUGE
CBLUE	Medical	Evaluate Chinese biomedical language understanding across multiple clinical and QA tasks	Accuracy, F1-score, and macro-average metrics across subtasks
PromptCBLUE	Medical	Evaluate LLMs on prompt-based generation across 16 Chinese medical NLP tasks	Accuracy, BLEU, and ROUGE scores
CMB	Medical	Evaluate comprehensive Chinese medical knowledge via exam-style QA and clinical diagnosis tasks	Accuracy, expert grading, and model-based evaluation
HuaTuo26M-test	Medical	Evaluate Chinese medical knowledge and QA ability using real-world clinical queries	Accuracy and relevance
CMExam	Medical	Evaluate LLMs on Chinese medical licensing exam QA with fine-grained clinical annotations	Accuracy, weighted F1 score, and expert-judged reasoning quality
MultiMedQA	Medical	Evaluate LLMs' clinical knowledge via multiple-choice and open-ended medical QA tasks	Expert-rated helpfulness, factuality, and safety
QiZhenGPT eval	Medical	Evaluate LLMs' ability to identify drug indications from natural language prompts	Expert-annotated correctness score
MedExQA	Medical	Evaluate LLMs' medical knowledge and explanation generation across underrepresented specialties	Explanation quality and expert-aligned LLM-judged relevance
JAMA and Medbullets	Medical	Evaluate LLMs' ability to answer and explain challenging clinical questions	Answer accuracy, explanation quality, and human-aligned reasoning assessment
MedXpertQA	Medical	Evaluate expert-level medical reasoning and multimodal clinical understanding across specialties	Answer accuracy, image-text reasoning performance, and expert-aligned scoring
MedJourney	Medical	Evaluate LLM performance across full clinical patient journey stages and tasks	Task-specific automatic metrics (e.g., accuracy) and human expert evaluations
MedAgentsBench	Medical	Evaluate complex multi-step clinical reasoning including diagnosis and treatment planning	Multi-aspect evaluation including correctness, efficiency, and human expert ratings
LongHealth	Medical (long-context)	Evaluate question answering over long-form clinical documents	Exact Match (EM), F1 score, and Long-Context QA accuracy
MedOdyssey	Medical (long-context)	Evaluate long-context understanding in the Medical domain	Task-specific exact match, ROUGE, and human preference scoring
HumanEval	Code	Evaluate code generation, algorithmic reasoning, and language understanding with functional correctness	Test-case execution accuracy (pass@k)
MBPP	Code	Evaluate basic Python code generation on crowdsourced programming tasks	Functional correctness via pass@k using automated test cases
APPS	Code	Evaluate coding challenge competence through real-world programming problems	Pass@k and exact match for functional correctness
DS-1000	Code	Evaluate data science code generation across real-world queries from 7 Python libraries	Functional correctness via automated test-based execution accuracy
CodeXGLUE	Code	Evaluate code understanding and generation across 9 tasks in 4 input-output types	Automatic metrics including BLEU, EM, F1, Accuracy, and MAP
CruxEval	Code	Evaluate code reasoning, understanding, and execution	pass@1 accuracy
ODEX	Code	Evaluate cross-lingual code generation from natural language queries in four languages	Functional correctness via execution-based evaluation
MTPB	Code	Evaluate multi-turn program synthesis	Functional correctness via pass@k on step-wise subprograms
ClassEval	Code	Evaluate class-level code generation from natural language descriptions in Python	pass@1, class completeness, dependency consistency, and error analysis
BigCodeBench	Code	Evaluate LLMs' ability to follow complex instructions and invoke diverse function calls	Pass@k, test case execution accuracy, branch coverage
HumanEvalPack	Code	Evaluate multilingual code generation, correction, and comment synthesis across six programming languages	Functional correctness via pass@k and task-specific auto metrics
BIRD	Code	Evaluate database-grounded text-to-SQL generation	Execution accuracy and exact match
RepoQA	Code (long-context)	Evaluate long-context code understanding in real-world software repositories	Exact match accuracy and retrieval-augmented correctness
LongCodeArena	Code (long-context)	Evaluate long-context code comprehension, generation, and editing across real-world tasks	Task-specific accuracy, exact match, and human evaluation
GSM8K	Math	Assess grade school math reasoning	Exact match accuracy
MATH	Math	Evaluate mathematical problem-solving ability on competition-level questions with step-by-step reasoning	Final answer accuracy and step-wise derivation correctness
MathOdyssey	Math	Evaluate LLMs' problem-solving capabilities on high school, university-level, and olympiad-level problems	Answer accuracy and performance across varying difficulty levels
MathBench	Math	Evaluate theoretical understanding and practical application of mathematical knowledge across five levels	Accuracy on theoretical and application problems
CHAMP	Math	Evaluate LLMs' fine-grained mathematical reasoning with concept and hint annotations on competition-level problems	Answer accuracy, reasoning path correctness
LILA	Math	Evaluate LLMs' mathematical reasoning across diverse formats and topics	Accuracy across 23 tasks
MiniF2F-v1	Math	Evaluate formal mathematical reasoning at Olympiad level	Proof accuracy on 488 problems
ProofNet	Math	Evaluate auto-formalization and formal proof generation in undergraduate-level mathematics	Formalization accuracy and formal proof success rate in Lean 3
AlphaGeometry	Math	Evaluate neuro-symbolic reasoning on olympiad-level Euclidean geometry theorems	Proof success rate, correctness, completeness, and human readability
MathVerse	Math	Evaluate MLLMs' capability to interpret and reason over visual math diagrams	Diagram-sensitive answer accuracy with fine-grained CoT reasoning score
We-Math	Math	Evaluate LMMs' visual mathematical reasoning with emphasis on knowledge acquisition and generalization	Four-dimensional diagnostic metrics
U-MATH	Math	Evaluate LLMs' open-ended problem-solving skills across university-level math with visual components	Solution correctness judged by LLMs with expert-verified F1 score
TabMWP	Math	Evaluate mathematical reasoning over textual and tabular data	Accuracy on QA and multiple-choice questions
MathHay	Math (long-context)	Evaluate long-context mathematical reasoning with multi-step dependencies	Accuracy, exact match, and reasoning chain consistency
LegalBench	Law	Evaluate legal reasoning across six types including rule application and interpretation	Task-specific accuracy, rule-consistency, and LLM-as-a-judge ratings
LexGLUE	Law	Evaluate legal language understanding across classification and QA tasks	Task-specific accuracy and F1 score (Auto)
LEXTREME	Law	Evaluate multilingual and multitask legal language understanding across 24 languages and 18 tasks	Task-specific accuracy, macro-F1, and classification metrics (Auto)
LawBench	Law	Evaluate Chinese legal LLMs across retention, understanding, and application dimensions via 20 tasks	Task-specific accuracy and F1
LAiW	Law	Evaluate Chinese legal LLMs across fundamental, basic, and advanced legal tasks	Task-specific accuracy and F1 across 13 assignments
LexEval	Law	Evaluate LLMs' Chinese legal understanding and reasoning using a taxonomy of legal cognitive abilities	Task-specific accuracy
CitaLaw	Law	Evaluate LLMs' ability to generate legally sound answers with appropriate citations using statutes and precedent cases	Syllogism-based alignment score, citation accuracy and legal consistency
LegalAgentBench	Law	Evaluate LLM agents' ability to solve complex real-world legal tasks	Task success rate and intermediate progress rate
SCALE	Law (long-context)	Evaluate LLMs' legal reasoning across languages, long documents, and multitask legal scenarios	Accuracy, F1, and code-based assessment for long-context legal tasks

data processing on devices with limited connectivity. FinQA (Chen et al., 2021b) and BizBench (Koncel-Kedziorski et al., 2023) benchmark multi-step numerical and quantitative reasoning, involving both tabular data and executable financial logic, useful in portfolio analysis, valuation, and budgeting applications. EconLogicQA (Quan & Liu, 2024) addresses sequential economic reasoning over multi-event scenarios. In the Chinese financial domain, FinEval (Zhang et al., 2023c), CFBenchmark (Lei et al., 2023), and BBT-Fin (Lu et al., 2023a) assess task-specific instruction following across topics such as taxation, accounting, and investment strategy. Hirano (Hirano, 2024) enables financial language understanding evaluation in Japanese, while MultiFin (Jørgensen et al., 2023) focuses on multilingual topic classification, useful in federated settings with cross-border clients or news sources.

(iii) Long-context financial reasoning. Many practical financial tasks involve extended documents such as annual reports, investor briefs, and regulatory filings. Long-context benchmarks are critical to evaluate whether FedLLM can perform document-level comprehension and multi-step derivation under memory and privacy constraints. DocFinQA (Reddy et al., 2024) simulates multi-step numerical reasoning over financial reports, measuring exact match, F1, and reasoning traceability. FinTextQA (Chen et al., 2024b) further tests open-ended QA over long textual contexts using BLEU and ROUGE for generation quality. These benchmarks reflect realistic federated scenarios, such as on-device due diligence support or local regulatory

interpretation, where global models must adapt to long-form content without accessing raw documents. A summary of these finance-specific evaluation benchmarks, including their evaluation objectives and corresponding metrics, is presented in Table 11.

5.2.3 Medical Domain Evaluation Benchmarks

Medical-specific evaluation benchmarks are essential for assessing the performance of LLMs in healthcare applications, where accuracy, safety, and domain-specific understanding are critical (Thirunavukarasu et al., 2023). In the context of FedLLM, these benchmarks are particularly important for evaluating models deployed in privacy-sensitive environments, such as hospital intranets, personal health monitoring devices, and clinical support systems, where patient data cannot be centralized due to regulatory and ethical constraints. To reflect the complexity of medical reasoning and language comprehension under such federated conditions, medical benchmarks cover a wide range of evaluation targets, including diagnostic reasoning, medical question answering, clinical document understanding, and guideline adherence.

(i) Medical knowledge and QA evaluation. This category focuses on assessing LLMs' general and specialized medical knowledge through structured question-answering tasks. CBLUE (Zhang et al., 2021) and PromptCBLUE[26] benchmark Chinese biomedical NLP tasks across information extraction, document classification, and QA, with the latter emphasizing prompt-based generation. CMB (Wang et al., 2023e), HuaTuo26M-test (Li et al., 2023d), and CMExam (Liu et al., 2023d) evaluate models on Chinese medical exam-style QA, clinical diagnosis tasks, and real-world query understanding. These benchmarks are well-suited for evaluating FedLLM tailored to localized clinical documentation and public health information. MultiMedQA (Singhal et al., 2023) serves as a high-quality English benchmark encompassing multiple-choice and open-ended medical QA, with expert-based evaluation of factuality, helpfulness, and safety—three pillars critical for deploying FedLLM in patient-facing applications. QiZhenGPT eval[27] provides an annotation-based evaluation of drug indication extraction from natural language prompts, useful for drug interaction checking at the edge. MedExQA (Kim et al., 2024) expands QA evaluation to underrepresented medical specialties, while JAMA and Medbullets (Chen et al., 2025) challenge models with high-difficulty US medical exam-style questions and demand strong explanation quality.

(ii) Clinical reasoning and care process modeling. In practice, many medical applications require multi-step diagnostic reasoning, care pathway modeling, and treatment planning, especially in multi-agent or longitudinal clinical scenarios. MedXpertQA (Zuo et al., 2025) tests both textual and multimodal reasoning, simulating specialist-level question answering across medical images and textual reports. MedJourney (Wu et al., 2024d) evaluates LLMs across the full patient care pipeline, from chief complaint triage to follow-up guidance, with both task-specific and human expert evaluation. MedAgentsBench (Tang et al., 2025) explicitly focuses on multi-turn clinical planning, assessing the model's ability to generate coherent and correct diagnostic and treatment steps over multiple interactions—an ideal setup for privacy-preserving agent-based FedLLM deployment.

(iii) Long-context clinical understanding. Federated medical applications often involve lengthy patient histories, radiology reports, or guideline documents. Benchmarks in this group evaluate the ability of LLMs to perform robust QA and inference over long-form inputs. LongHealth (Adams et al., 2024) focuses on QA over long clinical narratives, measuring exact match and F1 accuracy. MedOdyssey (Fan et al., 2024a) targets long-context understanding across clinical specialties, incorporating ROUGE, EM, and human preference scoring to assess consistency and informativeness over extended reasoning chains. A comprehensive summary of these medical evaluation benchmarks is provided in Table 11.

5.2.4 Code Domain Evaluation Benchmarks

Code-specific evaluation benchmarks are essential for assessing the ability of FedLLM to understand, generate, and reason over programming logic across multiple languages and task types (Jiang et al., 2024a). In federated settings, code-related applications are commonly deployed in heterogeneous environments, including personalized coding assistants, on-device tutoring systems, and localized software development workflows.

[26] https://github.com/michael-wzhu/PromptCBLUE
[27] https://github.com/CMKRG/QiZhenGPT/tree/main/data/eval

These scenarios place unique demands on LLMs: they must follow fine-grained instructions, ensure functional correctness, and maintain high reliability—all while operating under the resource constraints and privacy requirements that are characteristic of FedLLM deployment. Benchmarks in this domain typically evaluate code synthesis, completion, bug fixing, and multi-turn problem solving.

(i) Basic code generation and functional reasoning. This group of benchmarks focuses on evaluating models' ability to generate syntactically correct and semantically valid code from natural language instructions. HumanEval (Chen et al., 2021a) and MBPP (Austin et al., 2021) are foundational benchmarks evaluating one-shot Python function generation, judged via execution-based metrics like pass@k. APPS (Hendrycks et al., 2021a) and DS-1000 (Lai et al., 2023) extend this to more complex and real-world tasks, with DS-1000 focusing on data science scenarios across multiple Python libraries. CodeXGLUE (Lu et al., 2021) offers a broad suite of tasks—spanning code summarization, translation, and generation—making it suitable for assessing FedLLM that may specialize in different sub-tasks across clients. CruxEval (Gu et al., 2024a) further emphasizes logical reasoning and error-free execution in high-stakes contexts. ODEX (Wang et al., 2022c) introduces cross-lingual code generation, evaluating the model's ability to translate natural language into code in four different programming languages, a valuable benchmark for multilingual FedLLM deployment. These datasets are particularly useful for evaluating client-level specialization in federated setups, where users might work with domain-specific codebases or tools.

(ii) Multi-turn synthesis and structural understanding. Modern code development involves iterative and structural logic, making multi-turn and component-aware benchmarks crucial. MTPB (Nijkamp et al., 2022) focuses on multi-turn program synthesis, assessing the ability to generate partial, compositional subprograms in sequence. ClassEval (Du et al., 2023) evaluates whether LLMs can generate coherent Python classes, including method dependencies and variable interactions—reflecting real-world object-oriented programming needs. BigCodeBench (Zhuo et al., 2024) challenges models with complex, multi-functional instruction following, and rich code behavior evaluation, using criteria like test case accuracy and branch coverage. HumanEvalPack (Muennighoff et al., 2023) extends HumanEval's principles to six languages and multiple code-related subtasks, enabling multilingual federated evaluation. BIRD (Li et al., 2023f) adds a structured dimension by testing text-to-SQL generation for database querying, emphasizing schema-aware reasoning and executable correctness—an increasingly relevant task in enterprise AI agents and personal data querying under private data constraints. Such benchmarks are particularly relevant for FedLLM deployed in collaborative or enterprise environments, where partial programs must be incrementally refined by agents or users with limited compute.

(iii) Long-context code understanding and retrieval. Real-world software development often requires reasoning over extended codebases or repositories, posing a challenge for memory-constrained clients. RepoQA (Liu et al., 2024b) and LongCodeArena (Bogomolov et al., 2024) evaluate models' comprehension and retrieval-augmented reasoning over entire repositories or large code documents. They measure not only token-level accuracy but also structural coherence and retrieval efficacy. These benchmarks offer valuable insights into the performance of long-context-aware FedLLM on realistic software engineering tasks such as bug fixing, documentation generation, and legacy code comprehension—particularly under constraints imposed by limited local computational resources.

Together, these code-specific benchmarks provide a comprehensive foundation for assessing the capabilities and limitations of FedLLM in coding applications. A detailed overview of benchmark objectives and evaluation metrics is provided in Table 11.

5.2.5 Math Domain Evaluation Benchmarks

Mathematical reasoning tasks serve as a rigorous benchmark for evaluating the generalization, compositionality, and step-by-step problem-solving capabilities of LLMs (Ahn et al., 2024). In federated settings, math-specific evaluations are especially valuable for applications such as privacy-preserving intelligent tutoring systems, on-device educational tools, and personalized STEM learning assistants. These tasks present unique challenges for FedLLM, as they require precise multi-step reasoning, symbolic computation, and logical consistency—often under strict memory, computation, and communication constraints. As such, math

benchmarks are instrumental in assessing a model's ability to perform structured reasoning in resource-constrained and heterogeneous environments.

(i) Primary math reasoning across educational levels. Benchmarks in this group assess basic-to-advanced math problem solving and reasoning: GSM8K (Cobbe et al., 2021a) focuses on grade school arithmetic word problems, serving as a foundation for reasoning evaluation. MATH (Hendrycks et al., 2021d) extends this to competition-level questions in algebra, geometry, and calculus, emphasizing step-by-step derivation accuracy. MathOdyssey (Fang et al., 2024a) evaluates reasoning across high school, university, and Olympiad levels, providing a broad difficulty spectrum. MathBench (Liu et al., 2024a) systematically tests both theoretical understanding and practical application across five levels, reflecting multi-tier federated education scenarios. CHAMP (Mao et al., 2024) introduces concept and hint annotations, useful for federated tutoring agents that may require step-wise guidance or personalization for struggling learners. LILA (Mishra et al., 2022) further expands evaluation to 23 math task types, offering comprehensive insight into the model's versatility across mathematical formats.

(ii) Formal proof and symbolic reasoning. Mathematics often requires formal logic and symbolic structure, which tests a model's ability to generalize beyond pattern-matching: MiniF2F-v1 (Zheng et al., 2021) and ProofNet (Azerbayev et al., 2023) evaluate formal mathematical reasoning and proof generation, with the latter using Lean 3 as a backend for correctness verification. AlphaGeometry (Trinh et al., 2024) blends neural and symbolic reasoning for Euclidean geometry, a domain requiring precise spatial logic and theorem synthesis—especially relevant in expert-centric or research-level FedLLM deployment.

(iii) Visual and diagrammatic mathematical reasoning. Many real-world math problems include visual elements (graphs, tables, geometric diagrams), posing multi-modal reasoning challenges: MathVerse (Zhang et al., 2024f) and We-Math (Qiao et al., 2024) evaluate visual reasoning using diagram interpretation, requiring fine-grained attention to layout and symbolic grounding. U-MATH (Chernyshev et al., 2024) tests open-ended university-level questions involving visual cues, with LLM-assisted expert scoring. TabMWP (Lu et al., 2022b) focuses on text–table joint reasoning, simulating practical applications like report analysis or financial tutoring in federated agents.

(iv) Long-context mathematical reasoning. FedLLM deployed on real-world devices often faces scenarios where mathematical problems span multiple steps or documents: MathHay (Wang et al., 2024a) evaluates reasoning across extended input chains, testing memory retention and logic consistency in multi-hop math reasoning—an important benchmark for long-context capabilities in private and offline educational settings.

Together, these benchmarks form a robust and diverse suite for evaluating the mathematical competency of FedLLM under different input formats, difficulty levels, and reasoning demands. They are especially vital for personalized STEM learning assistants and edge-based automated math tutoring.

5.2.6 Legal Domain Evaluation Benchmarks

Legal AI applications impose stringent requirements on factual accuracy, contextual understanding, and logical consistency—making legal benchmarks particularly important for evaluating FedLLM designed for use in areas such as smart justice, personalized legal consultation, and privacy-preserving regulatory compliance. These benchmarks are designed to assess model capabilities across a range of legal reasoning tasks, including legal text comprehension, argument analysis, statutory interpretation, and multi-document synthesis. Moreover, they often adopt multilingual and long-context formats that reflect the real-world complexity and heterogeneity of legal documents—challenges that are amplified in federated settings (Chen et al., 2024e).

(i) Foundational and legal-specific reasoning. Benchmarks in this category assess general-purpose legal understanding, including statute interpretation, legal judgment tasks, and document comprehension: LegalBench (Guha et al., 2023) provides a suite of six legal reasoning tasks such as rule application and contract understanding, evaluating both correctness and rule-consistency. LexGLUE (Chalkidis et al., 2021) includes classic legal NLP tasks such as case classification, contract QA, and legal entailment, making it suitable for evaluating LLMs in general legal understanding scenarios. LEXTREME (Niklaus et al., 2023) expands this evaluation to 24 languages and 18 tasks, making it a critical multilingual benchmark for assessing FedLLM's cross-lingual legal proficiency in global regulatory contexts.

(ii) Chinese legal language understanding. Given the importance of regional legal systems, several benchmarks target Chinese legal capabilities, aligning well with privacy-sensitive applications deployed in Chinese jurisdictions: LawBench (Fei et al., 2023) evaluates Chinese legal LLMs across three cognitive levels—retention, understanding, and application—via 20 legal tasks. LAiW (Dai et al., 2023) offers a fine-grained assessment framework covering fundamental to advanced legal challenges. LexEval (Li et al., 2024b) structures evaluation around a taxonomy of legal cognitive abilities, including memory, reasoning, and application. These benchmarks support federated personalization and regional adaptation of legal agents.

(iii) Legal citation and generation with formal grounding. Some legal tasks require not only correct answers but also proper justification grounded in laws and precedents: CitaLaw (Zhang et al., 2024c) focuses on generating legally sound responses with accurate citations to statutes and precedent cases. It uses metrics like syllogism alignment and legal consistency, making it particularly relevant for FedLLM agents operating in jurisdictions with citation and traceability requirements. **(iv) Legal agents and dynamic task-solving.** FedLLM may increasingly support legal assistants or decision-support agents that require multi-step, tool-integrated reasoning: LegalAgentBench (Li et al., 2024a) provides a novel benchmark for evaluating LLM agents on complex, multi-turn legal scenarios. It incorporates intermediate progress tracking and task success scoring, which are useful for evaluating FedLLM performance in decentralized and asynchronous workflows.

(v) Long-context and multilingual legal reasoning. Legal documents are often lengthy, hierarchical, and span multiple statutes or precedents: SCALE (Rasiah et al., 2023) is designed to evaluate LLMs on long-context legal documents, legal multilingualism, and cross-document reasoning. Its inclusion of multi-task legal scenarios and code-level legal analysis makes it especially relevant for edge-deployed legal assistants in enterprise or government use cases. Together, these benchmarks provide a rich and diverse framework for evaluating FedLLM in the legal domain—where privacy, jurisdictional customization, long-context understanding, and reasoning fidelity are all critical to real-world deployment.

6 Application

6.1 FedLLM for Recommendation Systems

Recommendation systems are pivotal across domains such as e-commerce, content streaming, and personalized advertising (Ko et al., 2022). Traditional approaches often rely on centralized data collection, raising significant privacy concerns, particularly when handling sensitive user interactions and preferences. Federated fine-tuning offers a promising alternative by enabling collaborative learning across distributed clients while preserving data privacy.

Zhao et al. (2024a) propose FELLRec, a federated framework for LLM-based recommendation, to tackle the challenges of client performance imbalance and high resource costs. Specifically, FELLRec employs dynamic parameter aggregation and adaptive learning speeds to ensure balanced performance across clients. Additionally, it selectively retains sensitive LLM layers on the client side while offloading other layers to the server, effectively preserving privacy and optimizing resource usage. Similarly, Yuan et al. (2024c) introduce FELLAS, a federated sequential recommendation framework that leverages LLMs as external services to enhance sequential recommendation. FELLAS enriches item embeddings via LLM-assisted textual representation while ensuring privacy protection through d_x-privacy-compliant sequence perturbation. Beyond privacy protection, FL also facilitates reinforcement learning from human feedback for LLM-based recommendation systems. Wu et al. (2024c) propose FedBis and FedBiscuit, two frameworks designed to enable privacy-preserving federated RLHF. FedBis collaboratively trains a binary selector to filter sensitive preference data, while FedBiscuit clusters clients to train multiple selectors, ensuring better alignment with human preferences while maintaining privacy. Another framework, GPT-FedRec, proposed by Zeng et al. (2024a), integrates ChatGPT with a hybrid Retrieval-Augmented Generation (RAG) mechanism to address data sparsity, heterogeneity, and LLM-specific challenges in federated recommendation. GPT-FedRec employs hybrid retrieval techniques to extract user patterns and item features, then refines recommendations through LLM-generated prompts. By leveraging RAG, the framework effectively mitigates hallucination in LLM-generated content and enhances the overall recommendation quality.

In summary, as LLMs become increasingly integrated into recommendation systems, federated fine-tuning has emerged as a powerful method to fully leverage the capabilities of LLMs while ensuring user data privacy. These advancements demonstrate the potential of FedLLM to support high-quality, privacy-aware recommendation in real-world applications.

6.2 FedLLM for Biomedical Research

In the biomedical domain, direct data transmission and centralized model fine-tuning pose significant risks to user and patient privacy. Federated fine-tuning provides a privacy-preserving paradigm that enables collaborative model adaptation across decentralized medical datasets without exposing sensitive information. Ali et al. (2025) explore the use of various FL techniques to fine-tune time-series LLMs on electrocardiogram and impedance cardiography data, enabling privacy-preserving physiological signal analysis. Naseer & Nandakumar conduct a systematic investigation into the application of federated PEFT strategies for fine-tuning vision transformers in medical image classification tasks. Puppala et al. (2024) present a FL-based GPT chatbot designed for personalized healthcare information retrieval. The system aggregates and curates information from diverse sources while ensuring privacy and security through decentralized training. Users receive real-time, personalized insights via an intuitive interface, supported by advanced text parsing, metadata enrichment, and question-answering capabilities. This framework marks a key advancement in patient-centric AI applications.

Sarwar (2025) introduces FedMentalCare, a privacy-preserving framework that integrates FL and LoRA to fine-tune LLMs for mental health analysis. Their study explores the impact of client data volumes and model architectures (e.g., MobileBERT, MiniLM) in FL settings, ensuring scalability, data security, and computational efficiency. Liu et al. (2024c) propose FedFMS, which introduces federated foundation models for medical image segmentation. It addresses privacy challenges in medical imaging by enabling federated training without centralized data sharing. Wang et al. (2024d) introduce FEDKIM, a federated knowledge injection framework for scaling medical foundation models. It leverages lightweight local models to extract private knowledge and integrates it into a centralized model using an adaptive multitask multimodal mixture of experts module, enabling efficient cross-institution knowledge transfer. Dai et al. (2025) propose FedATA, a self-supervised FL framework for medical image segmentation, integrating masked self-distillation with adaptive attention to enhance pre-training and fine-tuning on unlabeled and limited-annotation data. Unlike traditional masked image modeling, FedATA uses latent representations as targets instead of pixels, improving feature learning. Additionally, its adaptive attention aggregation with personalized FL captures institution-specific representations, boosting model generalization and local fine-tuning performance.

In summary, LLMs have become increasingly influential in biomedical research. However, due to the highly sensitive nature of user data in this domain, federated fine-tuning has emerged as a pivotal approach for enabling large-scale AI applications in biomedicine and healthcare while ensuring data privacy and security.

6.3 FedLLM for Finance

The financial sector heavily relies on data-driven models for risk assessment, fraud detection, algorithmic trading, and personalized financial services. However, financial data is often highly sensitive, heavily regulated, and distributed across multiple institutions, making centralized model training infeasible due to privacy concerns and compliance constraints. Federated fine-tuning presents a promising solution by enabling collaborative learning across financial institutions without exposing raw data.

Ye et al. (2024b) introduce OpenFedLLM, a federated fine-tuning framework designed to train LLMs on decentralized private data while ensuring data privacy. In the financial domain, FL-tuned LLMs significantly outperform locally trained models and even surpass GPT-4, demonstrating the potential of FL to enhance LLM performance without compromising sensitive financial data. This study underscores the value of federated fine-tuning in leveraging distributed financial data to develop more accurate, robust, and privacy-preserving LLMs for financial applications. Shabani (2024) explore the use of FL for fine-tuning LLMs in finance, enhancing efficiency and privacy while addressing data scarcity and distribution challenges. Their findings show that FL achieves performance comparable to centralized fine-tuning with significantly lower computational costs and training time, making it ideal for resource-constrained environments. This ap-

proach preserves data privacy while enabling the development of more accurate and robust financial LLMs. Similarly, Zeng et al. (2024b) investigate fine-tuning financial LLMs using LoRA and deploying them on edge devices, demonstrating FL's potential to improve both model efficiency and performance in financial applications. Their study highlights significant gains in reasoning capabilities and cost-effectiveness, offering valuable insights into leveraging FL and LLMs for private and vertically specialized financial domains.

In summary, federated fine-tuning plays a crucial role in the financial sector by enabling collaborative model training across institutions while ensuring strict data privacy compliance. This approach allows financial organizations to leverage vast, decentralized datasets for fine-tuning, improving model accuracy without exposing sensitive financial information. As financial markets grow more complex and globally interconnected, FedLLM presents a scalable and secure pathway toward next-generation AI-driven financial infrastructure.

7 Open Challenges and Future Directions

Model Security of FedLLM. As federated fine-tuning gains momentum, ensuring model security has become a critical concern. In FedLLM, pre-trained models, whether proprietary or open-source, must be transmitted to distributed clients for local fine-tuning, inherently increasing the risk of intellectual property (IP) leakage and system vulnerabilities. Model security in this context involves two key aspects: protecting the IP of high-value models and ensuring the secure deployment of open-source models on edge devices.

First, the financial and strategic value of pre-trained LLMs makes IP protection in FedLLM deployment especially pressing. For example, training models like Gemini Ultra (Mesnard et al., 2024) and GPT-4 (OpenAI, 2023) is estimated to cost $191 million and $78 million, respectively. These models are typically developed by commercial entities under strict licensing and infrastructure control. However, in FedLLM settings, where the full model is often shared with clients in a white-box fashion, it becomes feasible for malicious participants to reverse-engineer or clone the model. This undermines the original developers' competitive advantage and deters participation from commercial model providers. Addressing this challenge requires the development of model watermarking (Pan et al., 2024a), encrypted model delivery, or inference-obfuscation protocols that allow clients to fine-tune and use the model without revealing sensitive architectural or parameter details.

Second, while open-source LLMs (e.g., DeepSeek (Bi et al., 2024), Qwen (Bai et al., 2023a)) are widely adopted in FedLLM due to their accessibility and flexibility, they present new vectors for security threats in federated deployments. In practice, most clients, especially those with limited machine learning or systems expertise, may lack the capabilities to deploy these models securely. For instance, frameworks like Ollama have been found to expose users to data leakage and unauthorized resource usage due to insecure default configurations (AIbase, 2025). In a federated setup, such vulnerabilities are amplified: a single compromised client can leak locally fine-tuned training data, or propagate adversarial backdoors to the global model. The consequences are particularly severe in sensitive domains like healthcare and finance, where breaches may result in the disclosure of protected health information (PHI) or proprietary trading strategies.

To mitigate these risks, future FedLLM research should prioritize the integration of secure model deployment practices into the federated fine-tuning pipeline. Techniques such as confidential computing for secure execution on edge devices, encrypted model delivery, and runtime access control should be incorporated to prevent unauthorized access and tampering. By embedding such security mechanisms into the FedLLM lifecycle, both commercial and open-source models can be safeguarded against misuse, thereby promoting broader adoption in high-stakes domains such as healthcare, finance, and critical infrastructure.

LLM and SLM Collaboration. A key future direction for FedLLM lies in enabling efficient collaboration between LLMs and small language models (SLMs) to address the performance–privacy–efficiency trade-offs inherent to federated settings. While LLMs offer superior reasoning and multi-modal capabilities, their large size and resource demands make them impractical for direct deployment on edge devices. Conversely, SLMs such as Gemini Nano (Team et al., 2023) and Phi-3 (Abdin et al., 2024) provide lightweight alternatives with better deployment efficiency but limited generalization and task transferability.

To reconcile these limitations, emerging FedLLM architectures can adopt a hybrid model paradigm: deploying SLMs at the edge for privacy-sensitive inference and lightweight tasks, while offloading complex reasoning

or orchestration to cloud-hosted LLMs. This collaborative strategy not only reduces the communication and computation burden at the client side but also enhances regulatory compliance by ensuring sensitive data never leaves the local device. Within this architecture, edge SLMs can perform initial text generation or instruction parsing, while LLMs handle tool selection, global coordination, or cross-domain alignment.

However, realizing this collaboration raises several open challenges in FedLLM: 1) minimizing latency and bandwidth overhead introduced by frequent SLM–LLM interactions; 2) preserving consistency and alignment between local SLM outputs and global LLM behavior; and 3) dynamically adapting task delegation strategies based on client heterogeneity, model confidence, and task complexity. Future FedLLM research should design decentralized orchestration protocols for efficient SLM–LLM coordination, and introduce privacy-preserving metadata exchange mechanisms to protect tool usage logs and inference traces. By enabling seamless SLM–LLM collaboration under federated environments, this hybrid architecture can unlock new levels of scalability, efficiency, and privacy in real-world FedLLM deployments.

Multi-Modal FedLLM. While existing FedLLM research has primarily focused on text-based tasks, emerging real-world applications increasingly require multi-modal capabilities, such as integrating visual, speech, and sensor modalities (Zhang et al., 2024a). Large multi-modal models (LMMs), including GPT-4V (Yang et al., 2023b) and LLaVA (Liu et al., 2023b), have demonstrated strong performance in centralized settings. However, extending these models to federated environments presents several unresolved challenges.

A primary difficulty lies in modality heterogeneity across clients (Peng et al., 2024; Ouyang et al., 2023). Devices may possess different input types—for example, some may only have textual data, while others may hold image–text pairs—leading to modality imbalance, where certain modalities are underrepresented in the training process (Fan et al., 2024b). This imbalance can degrade the model's generalization across modalities. Additionally, achieving cross-modal alignment (Gao et al., 2024c)—the model's ability to relate and reason across different modalities—becomes more difficult in federated setups, where paired data (e.g., image–caption pairs) cannot be shared centrally. Moreover, the high computational demands of LMMs further constrain their deployment and fine-tuning on edge devices with limited memory and processing capabilities (Liang et al., 2024; Jin et al., 2024).

To address these challenges, future research should develop modular and flexible tuning frameworks that allow each modality to be fine-tuned independently on client devices. This decoupling enables efficient local adaptation without requiring all modalities to be present on each client. Furthermore, modality-aware aggregation protocols—which weight client contributions based on modality type, data quality, and semantic consistency—can help mitigate imbalance and enhance global model performance. Promising directions also include federated cross-modal contrastive learning (Yu et al., 2023c), which can improve multi-modal alignment without requiring raw data exchange. Finally, to facilitate deployment in edge-centric applications such as smart healthcare, assistive robotics, and wearable systems, it is essential to design lightweight multi-modal architectures through techniques like knowledge distillation (Cai et al., 2024b) or dynamic subnetwork activation (Alam et al., 2022) that strike a balance between accuracy and resource efficiency.

Continual Learning in FedLLM. In dynamic federated environments, client data distributions and task objectives evolve over time, necessitating continual learning capabilities in FedLLM systems (Yoon et al., 2021; Wang et al., 2024e). Unlike traditional FL settings with fixed tasks and static datasets, real-world deployments require models to incrementally incorporate new knowledge without retraining from scratch. However, continual fine-tuning of LLMs introduces several unique challenges. The sheer size and overparameterization of LLMs make them prone to catastrophic forgetting during incremental updates (Huang et al., 2024a), especially when client participation is sparse or irregular. Moreover, repeated retraining across rounds is computationally expensive and often infeasible on edge devices with limited hardware resources.

To overcome these limitations, future research should investigate parameter-efficient continual learning strategies that enable local knowledge retention while supporting scalable global updates. Techniques such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), PEFT-based modular updates, and rehearsal methods using compressed memory buffers (Tiwari et al., 2022) are promising in mitigating forgetting without incurring prohibitive overhead. In addition, the design of lifelong personalization protocols—capable of adapting to each client's evolving task distribution under Non-IID and intermittent data

availability—remains an open research frontier. Developing such protocols requires balancing communication efficiency, privacy preservation, and model stability across heterogeneous learning trajectories.

Ultimately, enabling continual adaptation in FedLLM will be essential for long-term deployment in dynamic real-world scenarios, such as personalized healthcare, evolving legal compliance systems, or lifelong learning assistants. This calls for a shift from round-based static fine-tuning to streaming, task-aware federated adaptation frameworks that can incrementally evolve with users and environments.

Memory-Efficient FedLLM. Memory efficiency remains one of the most fundamental and restrictive bottlenecks in the deployment of FedLLM—often resulting in a binary feasibility condition: a client device either meets the memory requirements to participate in training or is entirely excluded (Wu et al., 2024j). Unlike other challenges such as communication overhead or data heterogeneity, which degrade performance but still permit participation, memory limitations can preclude participation altogether, particularly for edge devices with constrained hardware capabilities. Although PEFT techniques like LoRA substantially reduce the number of trainable parameters, they fall short of fully mitigating memory pressure. For example, fine-tuning LLaMA2-13B with LoRA still demands over 50 GB of peak memory—an order of magnitude beyond the capacity of most mobile phones, IoT devices, or embedded systems (Xu et al., 2024b; Tian et al., 2024a).

Addressing this limitation requires innovation at both the algorithmic and system levels. On the algorithmic side, emerging approaches such as dynamic layer-wise adaptation (Pan et al., 2024b), quantization-aware PEFT (e.g., QLoRA) (Dettmers et al., 2023), and structured model pruning (Wang et al., 2019b; Ma et al., 2023) offer promising pathways for reducing memory footprints during local fine-tuning. Complementing these are system-level solutions such as gradient checkpointing and accumulation (Gim & Ko, 2022), run-time memory-aware schedulers, and cloud-edge hybrid training architectures with selective computation offloading (Kumar et al., 2013), all of which aim to stretch the effective memory capacity of participating devices. To fully unlock the potential of FedLLM at scale, future research should explore co-designed frameworks that jointly optimize algorithmic efficiency and system-level deployment. Such holistic solutions can harmonize memory, computation, and communication trade-offs in real time—enabling resource-adaptive, privacy-preserving model customization across highly diverse client ecosystems.

Overcoming the memory barrier would expand the pool of eligible participants to include billions of low-memory edge devices that are currently sidelined from federated training. This not only enhances the inclusiveness, representativeness, and scalability of the FedLLM framework, but also opens the door to real-world deployments in settings such as home automation, wearables, and low-power industrial IoT platforms.

8 Conclusion

To the best of our knowledge, this is the *first* comprehensive survey dedicated to the federated fine-tuning of LLMs. We begin by introducing foundational background knowledge and identifying four core challenges through empirical analysis, which reveal the fundamental limitations that federated fine-tuning must overcome. We then review the latest relevant research papers, systematically organizing recent advances in parameter-efficient federated fine-tuning techniques. These approaches are categorized based on their methodologies, with detailed discussions on how each class of methods addresses the identified challenges. Furthermore, we present a comprehensive evaluation framework encompassing both fine-tuning datasets and evaluation benchmarks across different domains, offering a holistic framework for assessing FedLLM performance. Beyond methodological contributions, we highlight practical applications of FedLLM across domains. Finally, we outline promising future directions in this rapidly evolving field. While notable progress has been achieved, several pressing challenges remain open. Addressing these issues will be essential to unlocking the full potential of FedLLM and enabling its widespread deployment in practical, privacy-sensitive applications.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. Longhealth: A question answering benchmark with long clinical documents. *arXiv preprint arXiv:2401.14490*, 2024.

Khwaja Mutahir Ahmad, Qiao Liu, Abdullah Aman Khan, Yanglei Gan, and Changhao Huang. Prompt-enhanced federated learning for aspect-based sentiment analysis. In *2023 International Conference on Intelligent Communication and Computer Engineering (ICICCE)*, pp. 81–87. IEEE, 2023.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

AIbase. Security Risks! Ollama Large Model Tool Allegedly Contains Critical Vulnerabilities. *AI News*, Mar 2025. URL `https://www.aibase.com/news/15909`.

Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. *Advances in neural information processing systems*, 35:29677–29690, 2022.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models, 2025. URL `https://arxiv.org/abs/2502.17387`.

Mahad Ali, Curtis Lisle, Patrick W Moore, Tammer Barkouki, Brian J Kirkwood, and Laura J Brattain. Fine-tuning foundation models with federated learning for privacy preserving medical time series forecasting. *arXiv preprint arXiv:2502.09744*, 2025.

Loubna Ben Allal, Lewis Tunstall, Anton Lozhkov, Elie Bakouch, Guilherme Penedo, and Gabriel Martín Blázquez Hynek Kydlicek. Open r1: Evaluating llms on uncontaminated math competitions, 2025.

Omair Rashed Abdulwareth Almanifi, Chee-Onn Chow, Mau-Luen Tham, Joon Huang Chuah, and Jeevan Kanesan. Communication and computation efficiency in federated learning: A survey. *Internet of Things*, 22:100742, 2023.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H Ezzeldin, Qingfeng Liu, Kee-Bong Song, Mostafa El-Khamy, and Salman Avestimehr. Slora: Federated parameter efficient fine-tuning of language models. *arXiv preprint arXiv:2308.06522*, 2023.

Gaurav Bagwe, Xiaoyong Yuan, Miao Pan, and Lan Zhang. Fed-cprompt: Contrastive prompt for rehearsal-free federated continual learning. *arXiv preprint arXiv:2307.04869*, 2023.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27284–27293, 2024b.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023b.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023c.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*, 2023.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. *arXiv preprint arXiv:2411.14961*, 2024.

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, et al. Long code arena: a set of benchmarks for long-context code models. *arXiv preprint arXiv:2406.11612*, 2024.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term only fine-tuning of foundation models. 2022.

David Byrd and Antigoni Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the first ACM international conference on AI in finance*, pp. 1–9, 2020.

Yuji Byun and Jaeho Lee. Towards federated low-rank adaptation of language models with rank heterogeneity. *arXiv preprint arXiv:2406.17477*, 2024.

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Efficient federated learning for modern nlp. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pp. 1–16, 2023a.

Dongqi Cai, Yaozong Wu, Haitao Yuan, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Towards practical few-shot federated nlp. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pp. 42–48, 2023b.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024a.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024b.

Di Chai, Leye Wang, Liu Yang, Junxue Zhang, Kai Chen, and Qiang Yang. A survey for federated learning evaluations: Goals and measures. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*, 2021.

Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation, 2023.

Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor S Sheng, Huaiyu Dai, and Dejing Dou. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. *arXiv preprint arXiv:2310.15080*, 2023.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599, 2025.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11285–11293, 2024a.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*, 2024b.

Jingxue Chen, Hang Yan, Zhiyuan Liu, Min Zhang, Hu Xiong, and Shui Yu. When federated learning meets privacy-preserving computation. *ACM Computing Surveys*, 56(12):1–36, 2024c.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021a.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Shengchao Chen, Guodong Long, Tao Shen, and Jing Jiang. Prompt federated learning for weather forecasting: Toward foundation models on meteorological data. *arXiv preprint arXiv:2301.09152*, 2023a.

Shengchao Chen, Guodong Long, Tao Shen, Jing Jiang, and Chengqi Zhang. Federated prompt learning for weather foundation models on devices. *arXiv preprint arXiv:2305.14244*, 2023b.

Shuangyi Chen, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of foundation models via alternating minimization of lora. *arXiv preprint arXiv:2409.02346*, 2024d.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021b.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024e.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. U-math: A university-level benchmark for evaluating mathematical skills in llms. *arXiv preprint arXiv:2412.03205*, 2024.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous lora for federated finetuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12903–12913, 2024.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *arXiv preprint arXiv:2305.14938*, 2023.

Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53:5113–5155, 2020.

Igor Churin, Murat Apishev, Maria Tikhonova, Denis Shevelev, Aydar Bulatov, Yuri Kuratov, Sergej Averkiev, and Alena Fenogenova. Long input benchmark for russian analysis. *arXiv preprint arXiv:2408.02439*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021a.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021b.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.

Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. *arXiv preprint arXiv:2405.09771*, 2024.

Hui Dai, Dan Pechi, Xinyi Yang, Garvit Banga, and Raghav Mantri. Deniahl: In-context features influence llm needle-in-a-haystack abilities. *arXiv preprint arXiv:2411.19360*, 2024.

Jian Dai, Hao Wu, Huan Liu, Liheng Yu, Xing Hu, Xiao Liu, and Daoying Geng. Fedata: Adaptive attention aggregation for federated self-supervised medical image segmentation. *Neurocomputing*, 613:128691, 2025.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. Laiw: a chinese legal large language models benchmark. *arXiv preprint arXiv:2310.05620*, 2023.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532, 2020.

Wenlong Deng, Christos Thrampoulidis, and Xiaoxiao Li. Unlocking the potential of prompt-tuning in bridging generalized and personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6087–6097, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019a. doi: 10.18653/v1/n19-1423. URL `http://dx.doi.org/10.18653/v1/n19-1423`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019b. URL `https://arxiv.org/abs/1810.04805`.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023a.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023b.

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.

Chenhe Dong, Yuexiang Xie, Bolin Ding, Ying Shen, and Yaliang Li. Tunable soft prompts are messengers in federated learning. *arXiv preprint arXiv:2311.06805*, 2023a.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*, 2023b.

Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. *arXiv preprint arXiv:2308.01861*, 2023.

Yichao Du, Zhirui Zhang, Linan Yue, Xu Huang, Yuqing Zhang, Tong Xu, Linli Xu, and Enhong Chen. Communication-efficient personalized federated learning for speech-to-text tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10001–10005. IEEE, 2024.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. Medodyssey: A medical domain benchmark for long context evaluation up to 200k tokens. *arXiv preprint arXiv:2406.15019*, 2024a.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. Overcome modal bias in multi-modal federated learning via balanced modality selection. In *European Conference on Computer Vision*, pp. 178–195. Springer, 2024b.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*, 2024a.

Zihan Fang, Zheng Lin, Zhe Chen, Xianhao Chen, Yue Gao, and Yuguang Fang. Automated federated pipeline for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2404.06448*, 2024b.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.

Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8064–8073, 2023a.

Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4035–4045, 2023b.

Xiachong Feng, Xiaocheng Feng, Xiyuan Du, Min-Yen Kan, and Bing Qin. Adapter-based selective knowledge distillation for federated multi-domain meeting summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024a.

Yu Feng, Zhen Tian, Yifan Zhu, Zongfu Han, Haoran Luo, Guangwei Zhang, and Meina Song. Cp-prompt: Composition-based cross-modal prompting for domain-incremental continual learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2729–2738, 2024b.

Joseph G. Flowers. Finance-instruct-500k, 2025. URL `https://huggingface.co/datasets/Josephgflowers/Finance-Instruct-500k`.

Jörg Frohberg and Frank Binder. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. *arXiv preprint arXiv:2112.11941*, 2021.

Xingbo Fu, Binchi Zhang, Yushun Dong, Chen Chen, and Jundong Li. Federated graph machine learning: A survey of concepts, techniques, and applications. *ACM SIGKDD Explorations Newsletter*, 24(2):32–47, 2022.

Xingbo Fu, Zihan Chen, Yinhan He, Song Wang, Binchi Zhang, Chen Chen, and Jundong Li. Virtual nodes can help: Tackling distribution shifts in federated graph learning. *arXiv preprint arXiv:2412.19229*, 2024a.

Xingbo Fu, Zihan Chen, Binchi Zhang, Chen Chen, and Jundong Li. Federated graph learning with structure proxy alignment. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 827–838, 2024b.

Xingbo Fu, Song Wang, Yushun Dong, Binchi Zhang, Chen Chen, and Jundong Li. Federated graph learning with graphless clients. *arXiv preprint arXiv:2411.08374*, 2024c.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*, 2023.

Fei Gao, Yunfeng Zhao, Chao Qiu, Xiaofei Wang, Haipeng Yao, and Qinghua Hu. Cp 2 gfed: Cross-granular and personalized prompt-based green federated tuning for giant models. In *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*, pp. 1–10. IEEE, 2024a.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*, 2024b.

Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1860–1868, 2024c.

Sajjad Ghiasvand, Yifan Yang, Zhiyu Xue, Mahnoosh Alizadeh, Zheng Zhang, and Ramtin Pedarsani. Communication-efficient and tensorized federated fine-tuning of large language models. *arXiv preprint arXiv:2410.13097*, 2024.

In Gim and JeongGil Ko. Memory-efficient dnn training on mobile devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pp. 464–476, 2022.

Jack Goetz and Ambuj Tewari. Federated learning via synthetic data. *arXiv preprint arXiv:2008.04489*, 2020.

Shuai Gong, Chaoran Cui, Chunyun Zhang, Wenna Wang, Xiushan Nie, and Lei Zhu. Federated domain generalization via prompt learning and aggregation. *arXiv preprint arXiv:2411.10063*, 2024.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024a.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18099–18107, 2024b.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023a.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Lei Guo, Ziang Lu, Junliang Yu, Quoc Viet Hung Nguyen, and Hongzhi Yin. Prompt-enhanced federated content representation learning for cross-domain recommendation. In *Proceedings of the ACM on Web Conference 2024*, pp. 3139–3149, 2024a.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*, 2024b.

Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pp. 1364–1374, 2023b.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023c.

Tao Guo, Song Guo, and Junxiao Wang. Explore and cure: Unveiling sample effectiveness with context-aware federated prompt tuning. *IEEE Transactions on Mobile Computing*, 2024c.

Shaunak Halbe, James Seale Smith, Junjiao Tian, and Zsolt Kira. Hepco: Data-free heterogeneous prompt consolidation for continual federated learning. *arXiv preprint arXiv:2306.09970*, 2023.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18188–18196, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021a.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: an expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021c.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021d.

Masanori Hirano. Construction of a japanese financial benchmark for large language models. *arXiv preprint arXiv:2403.15062*, 2024.

Zhaoyi Joey Hou, Li Zhang, and Chris Callison-Burch. Choice-75: A dataset on decision branching in script learning. *arXiv preprint arXiv:2309.11737*, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*, 2024a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*, 2023a.

Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010, 2023b.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. Belle: Be everyone's large language model engine, 2023a.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*, 2023b.

Jingang Jiang, Xiangyang Liu, and Chenyou Fan. Low-parameter federated learning with large language models, 2023.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024a.

Yanna Jiang, Baihe Ma, Xu Wang, Guangsheng Yu, Ping Yu, Zhe Wang, Wei Ni, and Ren Ping Liu. Blockchained federated learning for internet of things: A comprehensive survey. *ACM Computing Surveys*, 56(10):1–37, 2024b.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. Multifin: A dataset for multilingual financial nlp. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 894–909, 2023.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A" novel" challenge for long-context language models. *arXiv preprint arXiv:2406.16264*, 2024.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

Gyunyeop Kim, Joon Yoo, and Sangwoo Kang. Efficient federated learning with pre-trained large language model using several adapter mechanisms. *Mathematics*, 11(21):4479, 2023a.

Yeachan Kim, Junho Kim, Wing-Lam Mok, Jun-Hyung Park, and SangKeun Lee. Client-customized adaptation for parameter-efficient federated learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1159–1172, 2023b.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533*, 2022.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*, 2023.

Jabin Koo, Minwoo Jang, and Jungseul Ok. Towards robust and efficient federated low-rank adaptation with heterogeneous clients. *arXiv preprint arXiv:2410.22815*, 2024.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Wei-Bin Kou, Qingfeng Lin, Ming Tang, Sheng Xu, Rongguang Ye, Yang Leng, Shuai Wang, Guofa Li, Zhenyu Chen, Guangxu Zhu, et al. pfedlvm: A large vision model (lvm)-driven and latent feature-based personalized federated learning framework in autonomous driving. *arXiv preprint arXiv:2405.04146*, 2024.

Wei-Bin Kou, Qingfeng Lin, Ming Tang, Rongguang Ye, Shuai Wang, Guangxu Zhu, and Yik-Chung Wu. Fast-convergent and communication-alleviated heterogeneous hierarchical federated learning in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning, 2023.

Karthik Kumar, Jibang Liu, Yung-Hsiang Lu, and Bharat Bhargava. A survey of computation offloading for mobile systems. *Mobile networks and Applications*, 18:129–140, 2013.

Kevin Kuo, Arian Raje, Kousik Rajesh, and Virginia Smith. Federated lora with sparse communication. *arXiv preprint arXiv:2406.05233*, 2024.

Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pp. 18319–18345. PMLR, 2023.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien MR Arnold, Vincent Perot, Siddharth Dalmia, et al. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*, 2024a.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. Ethic: Evaluating large language models on long-context tasks with high information coverage. *arXiv preprint arXiv:2410.16848*, 2024b.

Yang Lei, Jiangtong Li, Dawei Cheng, Zhijun Ding, and Changjun Jiang. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a.

Guanghao Li, Wansen Wu, Yan Sun, Li Shen, Baoyuan Wu, and Dacheng Tao. Visual prompt based personalized federated learning. *arXiv preprint arXiv:2303.08678*, 2023b.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. Legalagentbench: Evaluating llm agents in legal domain. *arXiv preprint arXiv:2412.17259*, 2024a.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*, 2024b.

Heju Li, Rui Wang, Jun Wu, and Wei Zhang. Federated edge learning via reconfigurable intelligent surface with one-bit quantization. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pp. 1055–1060. IEEE, 2022a.

Heju Li, Rui Wang, Wei Zhang, and Jun Wu. One bit aggregation for federated edge learning with reconfigurable intelligent surface: Analysis and optimization. *IEEE Transactions on Wireless Communications*, 22(2):872–888, 2022b.

Heju Li, Rui Wang, Jun Wu, Wei Zhang, and Ismael Soto. Reconfigurable intelligent surface empowered federated edge learning with statistical csi. *IEEE Transactions on Wireless Communications*, 23(6):6595–6608, 2023c.

Heju Li, Rui Wang, Mingyang Jiang, and Jianquan Liu. Star-ris empowered heterogeneous federated edge learning with flexible aggregation. *IEEE Internet of Things Journal*, 2025.

Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12151–12161, 2024c.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024d.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*, 2023d.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023e.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36:42330–42357, 2023f.

Shenghui Li, Fanghua Ye, Meng Fang, Jiaxu Zhao, Yun-Hin Chan, Edith C-H Ngai, and Thiemo Voigt. Synergizing foundation models and federated learning: A survey. *arXiv preprint arXiv:2406.12844*, 2024e.

Shitian Li, Chunlin Tian, Kahou Tam, Rui Ma, and Li Li. Breaking on-device training memory wall: A systematic survey. *arXiv preprint arXiv:2306.10388*, 2023g.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024f.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pp. 374–382, 2023h.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *arXiv preprint arXiv:2203.07814*, 2022c.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15 (6), 2023i.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

Li Lin, Yixiang Liu, Jiewei Wu, Pujin Cheng, Zhiyuan Cai, Kenneth KY Wong, and Xiaoying Tang. Fedlppa: Learning personalized prompt and aggregation for federated weakly-supervised medical image segmentation. *arXiv preprint arXiv:2402.17502*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Zihao Lin, Yan Sun, Yifan Shi, Xueqian Wang, Lifu Huang, Li Shen, and Dacheng Tao. Efficient federated prompt tuning for black-box large pre-trained models. *arXiv preprint arXiv:2310.03123*, 2023.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*, 2025.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, et al. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2305.10263*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024a.

Jiale Liu, Yu-Wei Zhan, Chong-Yu Zhang, Xin Luo, Zhen-Duo Chen, Yinwei Wei, and Xin-Shun Xu. Federated class-incremental learning with prompting. *arXiv preprint arXiv:2310.08948*, 2023c.

Jiawei Liu, Jia Le Tian, Vijay Daita, Yuxiang Wei, Yifeng Ding, Yuhan Katherine Wang, Jun Yang, and Lingming Zhang. Repoqa: Evaluating long context code understanding. *arXiv preprint arXiv:2406.06025*, 2024b.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452, 2023d.

Xiangyang Liu, Tianqi Pang, and Chenyou Fan. Federated prompting and chain-of-thought reasoning for improving llms answering. In *International Conference on Knowledge Science, Engineering and Management*, pp. 3–11. Springer, 2023e.

Yi Liu, Xiaohan Bi, Lei Li, Sishuo Chen, Wenkai Yang, and Xu Sun. Communication efficient federated learning for multilingual neural machine translation with adapter. *arXiv preprint arXiv:2305.12449*, 2023f.

Yuxi Liu, Guibo Luo, and Yuesheng Zhu. Fedfms: Exploring federated foundation models for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 283–293. Springer, 2024c.

Renze Lou, Kai Zhang, and Wenpeng Yin. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095, 2024.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*, 2023a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022a.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022b.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*, 2021.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023b.

Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. *arXiv preprint arXiv:2410.10114*, 2024.

Jialiang Ma, Chunlin Tian, Li Li, and Chengzhong Xu. Fedmg: A federated multi-global optimization framework for autonomous driving control. In *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*, pp. 1–10. IEEE, 2024a.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Yuting Ma, Lechao Cheng, Yaxiong Wang, Zhun Zhong, Xiaohua Xu, and Meng Wang. Fedhpl: Efficient heterogeneous federated learning with prompt tuning and logit distillation. *arXiv preprint arXiv:2405.17267*, 2024b.

Seiji Maekawa, Hayate Iso, and Nikita Bhutani. Holistic reasoning with long-context lms: A benchmark for database operations on massive textual data. *arXiv preprint arXiv:2410.11996*, 2024.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pp. 1941–1942, 2018.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

Yujun Mao, Yoon Kim, and Yilun Zhou. Champ: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities. *arXiv preprint arXiv:2401.06961*, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295, 2024. doi: 10.48550/ARXIV.2403.08295. URL `https://doi.org/10.48550/arXiv.2403.08295`.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*, 2022.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math, 2024.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro Von Werra, and Shayne Longpre. Octopack: Instruction tuning code large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Muzammal Naseer and Karthik Nandakumar. Probing the efficacy of federated parameter-efficient fine-tuning of vision transformers for medical image. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2024 Workshops: ISIC 2024, iMIMIC 2024, EARTH 2024, DeCaF 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6–10, 2024, Proceedings*, pp. 236. Springer Nature.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3):1–37, 2022.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, et al. Cfinbench: A comprehensive chinese financial benchmark for large language models. *arXiv preprint arXiv:2407.02301*, 2024.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*, 2023.

Zhiyuan Ning, Chunlin Tian, Meng Xiao, Wei Fan, Pengyang Wang, Li Li, Pengfei Wang, and Yuanchun Zhou. Fedgcs: A generative framework for efficient client selection in federated learning via gradient-based optimization. *arXiv preprint arXiv:2405.06312*, 2024.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL `https://doi.org/10.48550/arXiv.2303.08774`.

OpenAI. Hello GPT-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`.

Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pp. 530–543, 2023.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*, 2024a.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. Lisa: layerwise importance sampling for memory-efficient large language model fine-tuning. *Advances in Neural Information Processing Systems*, 37:57018–57049, 2024b.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Codeforces cots. `https://huggingface.co/datasets/open-r1/codeforces-cots`, 2025.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Yuanzhe Peng, Jieming Bian, and Jie Xu. Fedmm: Federated multi-modal learning with modality heterogeneity in computational pathology. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1696–1700. IEEE, 2024.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020.

Sai Puppala, Ismail Hossain, Md Jahangir Alam, and Sajedul Talukder. Scan: A healthcare personalized chatbot with federated learning based gpt. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1945–1951. IEEE, 2024.

Jiaxing Qi, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. Fdlora: Personalized federated learning of large language model via dual lora tuning. *arXiv preprint arXiv:2406.07925*, 2024a.

Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. Long2rag: Evaluating long-context & long-form retrieval-augmented generation with key point recall. *arXiv preprint arXiv:2410.23000*, 2024b.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023a.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. *arXiv preprint arXiv:2312.06353*, 2023b.

Zhen Qin, Zhaomin Wu, Bingsheng He, and Shuiguang Deng. Federated data-efficient instruction tuning for large language models. *arXiv preprint arXiv:2410.10926*, 2024.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Text-driven prompt generation for vision-language models in federated learning. *arXiv preprint arXiv:2310.06123*, 2023.

Yinzhu Quan and Zefang Liu. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning. *arXiv preprint arXiv:2405.07938*, 2024.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. Scale: Scaling up the complexity for advanced language model evaluation. *arXiv preprint arXiv:2306.09237*, 2023.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*, 2024.

Chao Ren, Han Yu, Hongyi Peng, Xiaoli Tang, Anran Li, Yulan Gao, Alysa Ziying Tan, Bo Zhao, Xiaoxiao Li, Zengxiang Li, et al. Advances and open challenges in federated learning with foundation models. *arXiv preprint arXiv:2404.15381*, 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL `https://arxiv.org/abs/1910.01108`.

SM Sarwar. Fedmentalcare: Towards privacy-preserving fine-tuned llms to analyze mental health status using federated learning framework. *arXiv preprint arXiv:2503.05786*, 2025.

Naser Shabani. Harnessing federated learning for llm fine-tuning: A distributed approach, 2024.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv e-prints*, pp. arXiv–2302, 2023.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*, 2022.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*, 2023.

shareAI. Sharegpt-chinese-english-90k bilingual human-machine qa dataset. `https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k`, 2023.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pp. 595–597, 2022.

Ming Shen. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*, 2024.

Jiyun Shin, Jinhyun Ahn, Honggu Kang, and Joonhyuk Kang. Fedsplitx: Federated split learning for computationally-constrained heterogeneous clients. *arXiv preprint arXiv:2310.14579*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Shangchao Su, Bin Li, and Xiangyang Xue. Fedra: A random allocation strategy for federated tuning to unleash the power of heterogeneous clients. *arXiv preprint arXiv:2311.11227*, 2023.

Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15117–15125, 2024.

Guangyu Sun, Umar Khalid, Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, and Chen Chen. Conquering the communication constraints to enable large pre-trained models in federated learning. *arXiv preprint arXiv:2210.01708*, 2022.

Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yiran Chen, and Holger R Roth. Fedbpt: Efficient federated black-box prompt tuning for large language models. *arXiv preprint arXiv:2310.01467*, 2023.

Yuchang Sun, Yuexiang Xie, Bolin Ding, Yaliang Li, and Jun Zhang. Exploring selective layer fine-tuning in federated learning. *arXiv preprint arXiv:2408.15600*, 2024.

Kahou Tam, Li Li, Bo Han, Chengzhong Xu, and Huazhu Fu. Federated noisy client learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023a.

Kahou Tam, Li Li, Yan Zhao, and Chengzhong Xu. Fedcoop: Cooperative federated learning for noisy labels. In *ECAI 2023*, pp. 2298–2306. IOS Press, 2023b.

Kahou Tam, Kewei Xu, Li Li, and Huazhu Fu. Towards federated domain unlearning: Verification methodologies and challenges. *arXiv preprint arXiv:2406.03078*, 2024.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*, 2025.

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, Jianye Hou, and Min Zhang. L-citeeval: Do long-context models truly leverage context for responding? *arXiv preprint arXiv:2410.02115*, 2024a.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. Mathscale: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*, 2024b.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

Omid Tavallaie and Niousha Nazemi[1]. Rbla: Rank-based-lora-aggregation for fine-tuning heterogeneous models. In *Web Services–ICWS 2024: 31st International Conference, Held as Part of the Services Conference Federation, SCF 2024, Bangkok, Thailand, November 16–19, 2024, Proceedings*, pp. 47. Springer Nature.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. Splitfed: When federated learning meets split learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8485–8493, 2022.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Chunlin Tian, Li Li, Zhan Shi, Jun Wang, and ChengZhong Xu. Harmony: Heterogeneity-aware hierarchical management for federated learning system. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 631–645. IEEE, 2022a.

Chunlin Tian, Zhan Shi, and Li Li. Learn to select: Efficient cross-device federated learning via reinforcement learning. In Krystal Maughan, Rosanne Liu, and Thomas F. Burns (eds.), *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=wecTsVkrjit`.

Chunlin Tian, Li Li, Kahou Tam, Yebo Wu, and Cheng-Zhong Xu. Breaking the memory wall for heterogeneous federated learning via model splitting. *IEEE Transactions on Parallel and Distributed Systems*, 2024a.

Chunlin Tian, Xinpeng Qin, and Li Li. Greenllm: Towards efficient large language model via energy-aware pruning. In *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*, pp. 1–2. IEEE, 2024b.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv preprint arXiv:2404.19245*, 2024c.

Chunlin Tian, Zhan Shi, Li Li, Cheng-zhong Xu, et al. Ranking-based client imitation selection for efficient federated learning. In *Forty-first International Conference on Machine Learning*, 2024d.

Yuanyishu Tian, Yao Wan, Lingjuan Lyu, Dezhong Yao, Hai Jin, and Lichao Sun. Fedbert: When federated learning meets pre-training. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4): 1–26, 2022b.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 99–108, 2022.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv: Arxiv-2402.10176*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023a.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. Progress in machine translation. *Engineering*, 18:143–153, 2022a.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020a.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623, 2020b.

Jie Wang, Yebo Wu, Erwu Liu, Xiaolong Wu, Xinyu Qu, Yuanzhe Geng, and Hanfu Zhang. Fedins2: A federated-edge-learning-based inertial navigation system with segment fusion. *IEEE Internet of Things Journal*, 11(2):3653–3661, 2023b.

Jie Wang, Xiaolong Wu, Jindong Tian, Erwu Liu, Yebo Wu, Rucong Lai, and Yong Tian. Indoor localization fusing inertial navigation with monocular depth estimation in federated learning framework with data heterogeneity. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Lei Wang, Shan Dong, Yuhui Xu, Hanze Dong, Yalu Wang, Amrita Saha, Ee-Peng Lim, Caiming Xiong, and Doyen Sahoo. Mathhay: An automated benchmark for long-context mathematical reasoning in llms. *arXiv preprint arXiv:2410.04698*, 2024a.

Lun Wang, Yang Xu, Hongli Xu, Zhida Jiang, Min Chen, Wuyang Zhang, and Chen Qian. Bose: Block-wise federated learning in heterogeneous edge computing. *IEEE/ACM Transactions on Networking*, 2023c.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*, 2024b.

Rui Wang, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Handong Zhao, Junda Wu, Subrata Mitra, Lina Yao, and Ricardo Henao. Personalized federated learning for text classification with gradient-free prompt tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4597–4612, 2024c.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications*, 37(6):1205–1221, 2019a.

Xiaochen Wang, Jiaqi Wang, Houping Xiao, Jinghui Chen, and Fenglong Ma. Fedkim: Adaptive federated knowledge injection into medical foundation models. *arXiv preprint arXiv:2408.10276*, 2024d.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023d.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*, 2023e.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023f.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Zhiruo Wang, Shuyan Zhou, Daniel Fried, and Graham Neubig. Execution-based evaluation for open-domain code generation. *arXiv preprint arXiv:2212.10481*, 2022c.

Zi Wang, Fei Wu, Feng Yu, Yurui Zhou, Jia Hu, and Geyong Min. Federated continual learning for edge-ai: A comprehensive survey. *arXiv preprint arXiv:2411.13740*, 2024e.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. Structured pruning of large language models. *arXiv preprint arXiv:1910.04732*, 2019b.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *arXiv preprint arXiv:2409.05976*, 2024f.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

Alex Watson, Yev Meyer, Maarten Van Segbroeck, Matthew Grossman, Sami Torbey, Piotr Mlocek, and Johnny Greco. Synthetic-PII-Financial-Documents-North-America: A synthetic dataset for training language models to label and detect pii in domain specific formats, June 2024. URL `https://huggingface.co/datasets/gretelai/synthetic_pii_finance_multilingual`.

Guoyizhe Wei, Feng Wang, Anshul Shah, and Rama Chellappa. Dual prompt tuning for domain-aware federated learning. *arXiv preprint arXiv:2310.03103*, 2023.

Pei-Yau Weng, Minh Hoang, Lam M Nguyen, My T Thai, Tsui-Wei Weng, and Trong Nghia Hoang. Probabilistic federated prompt-tuning with non-iid and imbalanced data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Herbert Woisetschläger, Alexander Isenko, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. A survey on efficient federated learning methods for foundation model training. *arXiv preprint arXiv:2401.04472*, 2024.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024a.

Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. Fedbiot: Llm local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3345–3355, 2024b.

Feijie Wu, Xiaoze Liu, Haoyu Wang, Xingchen Wang, and Jing Gao. On the client preference of llm fine-tuning in federated learning. *arXiv preprint arXiv:2407.03038*, 2024c.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Jie Yang, et al. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646, 2024d.

Xiaodong Wu, Minhao Wang, Yichen Liu, Xiaoming Shi, He Yan, Xiangju Lu, Junmin Zhu, and Wei Zhang. Lifbench: Evaluating the instruction following performance and stability of large language models in long-context scenarios. *arXiv preprint arXiv:2411.07037*, 2024e.

Xinghao Wu, Xuefeng Liu, Jianwei Niu, Haolin Wang, Shaojie Tang, and Guogang Zhu. Fedlora: When personalized federated learning meets low-rank adaptation. 2024f.

Xinghao Wu, Jianwei Niu, Xuefeng Liu, Mingjia Shi, Guogang Zhu, and Shaojie Tang. Tackling feature-classifier mismatch in federated learning via prompt-driven feature transformation. *arXiv preprint arXiv:2407.16139*, 2024g.

Yebo Wu, Li Li, Mang Ye, Chunlin Tian, KaHou Tam, He Sun, and Cheng-zhong Xu. Honey: Harmonizing progressive federated learning via elastic synergy across different training blocks.

Yebo Wu, Li Li, Chunlin Tian, Tao Chang, Chi Lin, Cong Wang, and Cheng-Zhong Xu. Heterogeneity-aware memory efficient federated learning via progressive layer freezing. In *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*, pp. 1–10. IEEE, 2024h.

Yebo Wu, Li Li, Chunlin Tian, Dubing Chen, and Chengzhong Xu. Neulite: Memory-efficient federated learning via elastic progressive training. *arXiv preprint arXiv:2408.10826*, 2024i.

Yebo Wu, Li Li, Chunlin Tian, and Chengzhong Xu. Breaking the memory wall for heterogeneous federated learning with progressive training. *arXiv preprint arXiv:2404.13349*, 2024j.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023a.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023b.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*, 2023c.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*, 2023d.

Liang Xu, Lei Zhu, Yaotong Wu, and Hang Xue. Superclue-fin: Graded fine-grained analysis of chinese llms on diverse financial tasks and applications. *arXiv preprint arXiv:2404.19063*, 2024a.

Mengwei Xu, Yaozong Wu, Dongqi Cai, Xiang Li, and Shangguang Wang. Federated fine-tuning of billion-sized language models across mobile devices. *arXiv preprint arXiv:2308.13894*, 2023e.

Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. {FwdLLM}: Efficient federated finetuning of large language models with perturbed inferences. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pp. 579–596, 2024b.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024c.

Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, et al. Detectiveqa: Evaluating long-context reasoning on detective novels. *arXiv preprint arXiv:2409.02465*, 2024d.

Yuxuan Yan, Qianqian Yang, Shunpu Tang, and Zhiguo Shi. Federa: Efficient fine-tuning of language models in federated learning leveraging weight decomposition. *arXiv preprint arXiv:2404.18848*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fu-En Yang, Chien-Yi Wang, and Yu-Chiang Frank Wang. Efficient model personalization in federated learning via client-specific prompt generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19159–19168, 2023a.

Jianxin Yang. Firefly: Chinese conversational large language models. `https://github.com/yangjianxin1/Firefly`, 2023.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*, 2022.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19368–19376, 2024a.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. *arXiv preprint arXiv:2403.19211*, 2024b.

Yuning Yang, Xiaohong Liu, Tianrun Gao, Xiaodong Xu, and Guangyu Wang. Sa-fedlora: Adaptive parameter allocation for efficient federated learning with lora tuning. *arXiv preprint arXiv:2405.09394*, 2024c.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1, 2023b.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Dezhong Yao, Wanning Pan, Michael J O'Neill, Yutong Dai, Yao Wan, Hai Jin, and Lichao Sun. Fedhm: Efficient federated learning for heterogeneous models via low-rank factorization. *arXiv preprint arXiv:2111.14655*, 2021.

Yuhang Yao, Jianyi Zhang, Junda Wu, Chengkai Huang, Yu Xia, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Ang Li, et al. Federated large language models: Current progress and future directions. *arXiv preprint arXiv:2409.15723*, 2024.

Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

Rongguang Ye, Wei-Bin Kou, and Ming Tang. Prafﬂ: A preference-aware scheme in fair federated learning. *arXiv preprint arXiv:2404.08973*, 2024a.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6137–6147, 2024b.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Jingwei Yi, Fangzhao Wu, Huishuai Zhang, Bin Zhu, Tao Qi, Guangzhong Sun, and Xing Xie. Robust quantity-aware aggregation for federated learning. *arXiv preprint arXiv:2205.10848*, 2022.

Liping Yi, Han Yu, Gang Wang, and Xiaoguang Liu. Fedlora: Model-heterogeneous personalized federated learning with lora tuning. *arXiv preprint arXiv:2310.13283*, 2023.

Xiaoyang Yi, Jian Zhang, Jing Chen, Yuru Bao, and Lingkai Xing. Fedfld: Heterogeneous federated learning via forget-less distillation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Xunjian Yin, Baizhou Huang, and Xiaojun Wan. Alcuna: Large language models meet new knowledge. *arXiv preprint arXiv:2310.14820*, 2023.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, pp. 12073–12086. PMLR, 2021.

Hao Yu, Xin Yang, Xin Gao, Yan Kang, Hao Wang, Junbo Zhang, and Tianrui Li. Personalized federated continual learning via multi-granularity prompt. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4023–4034, 2024.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*, 2023a.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023b.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023c.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL `https://arxiv.org/abs/2503.14476`.

Sixing Yu, J Pablo Muñoz, and Ali Jannesari. Bridging the gap between foundation models and heterogeneous federated learning. *arXiv preprint arXiv:2310.00247*, 2023d.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38, 2022.

Liangqi Yuan, Ziran Wang, Lichao Sun, S Yu Philip, and Christopher G Brinton. Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, 2024a.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, et al. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. *arXiv preprint arXiv:2402.05136*, 2024b.

Wei Yuan, Chaoqun Yang, Guanhua Ye, Tong Chen, Nguyen Quoc Viet Hung, and Hongzhi Yin. Fellas: Enhancing federated sequential recommendation with llm as external services. *ACM Transactions on Information Systems*, 2024c.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023a.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023b.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023c.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 9241–9250, 2020.

Hui Zeng. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*, 2023.

Huimin Zeng, Zhenrui Yue, Qian Jiang, and Dong Wang. Federated recommendation via hybrid retrieval augmented generation. *arXiv preprint arXiv:2403.04256*, 2024a.

Juntao Zeng, Bo Chen, Yuandan Deng, Weiqin Chen, Yanlin Mao, and Jiawei Li. Fine-tuning of financial large language model and application at edge device. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 42–47, 2024b.

Shichen Zhan, Yebo Wu, Chunlin Tian, Yan Zhao, and Li Li. Heterogeneity-aware coordination for federated learning via stitching pre-trained blocks. In *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*, pp. 1–10. IEEE, 2024.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM Symposium at IJCAI 2023*, 2023a.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024a.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919. IEEE, 2024b.

Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. Citalaw: Enhancing llm with citations in legal domain. *arXiv preprint arXiv:2412.14556*, 2024c.

Lei Zhang, Yunshui Li, Ziqiang Liu, Junhao Liu, Longze Chen, Run Luo, Min Yang, et al. Marathon: A race through the realm of long context with large language models. *arXiv preprint arXiv:2312.09542*, 2023b.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*, 2023c.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*, 2021.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024d.

Pengyu Zhang, Yingbo Zhou, Ming Hu, Junxian Feng, Jiawen Weng, and Mingsong Chen. Personalized federated instruction tuning via neural architecture search. *arXiv preprint arXiv:2402.16919*, 2024e.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024f.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*, 2023d.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023e.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505, 2023f.

Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiguang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Meikang Qiu, Sophia Ananiadou, Min Peng, et al. Dólares or dollars? unraveling the bilingual prowess of financial llms between spanish and english. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6236–6246, 2024g.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*, 2023g.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023h.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. ınftybench: Extending long context evaluation beyond 100k tokens. In *ACL (1)*, 2024h.

Yicheng Zhang, Zhen Qin, Zhaomin Wu, and Shuiguang Deng. Personalized federated fine-tuning for llms via data-driven heterogeneous model architectures. *arXiv preprint arXiv:2411.19128*, 2024i.

Yifei Zhang, Dun Zeng, Jinglong Luo, Xinyu Fu, Guanzhong Chen, Zenglin Xu, and Irwin King. A survey of trustworthy federated learning: Issues, solutions, and challenges. *ACM Transactions on Intelligent Systems and Technology*, 15(6):1–47, 2024j.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. *arXiv preprint arXiv:2406.02472*, 2024k.

Zikai Zhang, Jiahao Xu, Ping Liu, and Rui Hu. Fed-pilot: Optimizing lora assignment for efficient federated foundation model fine-tuning. *arXiv preprint arXiv:2410.10200*, 2024l.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Jujia Zhao, Wenjie Wang, Chen Xu, Zhaochun Ren, See-Kiong Ng, and Tat-Seng Chua. Llm-based federated recommendation. *arXiv preprint arXiv:2402.09959*, 2024a.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*, 2024b.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023b.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023a.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*, 2019.

Zibin Zheng, Kaiwen Ning, Yanlin Wang, Jingwen Zhang, Dewu Zheng, Mingxi Ye, and Jiachi Chen. A survey of large language models for code: Evolution, benchmarking, and future trends. *arXiv preprint arXiv:2311.10372*, 2023b.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv e-prints*, pp. arXiv–2306, 2023a.

Wei Zhu, Xiaoling Wang, and Longyue Wang. Chatmed: A chinese medical large language model. *Retrieved September*, 18:2023, 2023b.

WY Wei Zhu and Xiaoling Wang. Shennong-tcm: A traditional chinese medicine large language model. *GitHub*, 2023.

Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*, 2023.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*, 2024.

Kaijian Zou, Muhammad Khalifa, and Lu Wang. Retrieval or global context understanding? on many-shot in-context learning for long-context evaluation. *arXiv preprint arXiv:2411.07130*, 2024.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.