Theory of Mind Needs Privacy: A Collaborative SLM–LLM Framework

Yiming Luo[1] , **Chunlin Tian**[1] , **Xuyang Wei**[2]
[1]University of Macau
[2]University of Electronic Science and Technology of China
{dc22922, yc27402}@um.edu.mo, 7yun1uxdjj@gmail.com

Abstract

Large language models (LLMs) have significantly expanded the capabilities of AI systems, powering applications across domains such as finance, healthcare, and mathematical reasoning. In high-stakes contexts like psychological consultation, LLMs offer the potential to serve as intelligent assistants by providing fluent, contextually appropriate responses. However, this growing utility is accompanied by serious concerns. On one hand, user interactions with LLMs often involve the disclosure of highly sensitive personal data, raising critical issues of privacy and trust. On the other hand, LLMs are prone to hallucination—generating plausible but factually incorrect responses—which poses significant risks in domains that demand factual consistency and nuanced understanding. To address these challenges, prior work has proposed supervised fine-tuning (SFT) and retrieval-augmented generation (RAG) as two major solutions. While SFT adapts models to domain-specific knowledge, it is computationally expensive and often leads to overfitting or superficial memorization. RAG, in contrast, enables external knowledge retrieval but requires data centralization and exposes user queries to cloud-based systems, which undermines privacy guarantees. We advocate a hybrid paradigm that combines locally fine-tuned Small Language Models (SLMs) with a general-purpose LLM. This design ensures privacy, reduces hallucinations, and produces more contextually grounded responses through collaborative inference.

1 Introduction

Understanding human cognition and replicating human-like reasoning remains a long-standing challenge in artificial intelligence. Although LLMs have achieved remarkable progress in language generation and task execution, they still struggle with Theory of Mind (ToM)-related tasks that require deep reasoning, contextual understanding, and mental state modeling [Mirzadeh *et al.*, 2024]. Recent studies show that LLMs often rely on rote memorization rather than genuine generalization [Chu *et al.*, 2025], limiting their reliability in tasks that demand abstract inference, such as psychological counseling or multi-turn dialog reasoning.

This gap in reasoning ability restricts their effectiveness in scenarios that require deep inference and abstract thinking. As a result, LLMs often underperform on tasks that demand strong reasoning capabilities, and their performance in complex, multiturn interactions, such as those required in psychological counseling or other high-stakes dialog applications, remains suboptimal [Laban *et al.*, 2025; Kwan *et al.*, 2024; Chandra *et al.*, 2025]. For example, studies show that LLMs tend to make early incorrect assumptions in multi-turn conversations and struggle to recover, leading to cascading errors unlike what we see in single-turn settings. Furthermore, in multi-turn mental health counseling scenarios, LLMs score poorly on patient-centric communication and diagnostic reasoning, with performance degrading as turns progress. Though we have some advanced technologies, like Chain-of-Thought (CoT), the performance would still be poor. Another widely observed issue is the high incidence of hallucinations when LLMs are applied to reasoning or comprehension-oriented tasks [Huang *et al.*, 2025; Yao *et al.*, 2025]. These models can generate outputs that appear fluent and plausible yet are factually incorrect or misleading. Such hallucinations pose significant risks, especially in human–AI interactions involving vulnerable users. For instance, in conversations with children, an LLM that produces hallucinated responses may unintentionally introduce incorrect concepts, leading to substantial misunderstandings and cognitive distortions. This raises serious concerns about the reliability and safety of deploying LLMs in educational or developmental contexts where accuracy and trust are paramount.

To address these critical challenges, several solutions have been proposed within the research community. One commonly adopted approach is SFT, where a model is trained on domain-specific datasets to enhance its performance on targeted tasks. Another widely studied method is RAG, which equips an LLM with access to an external knowledge base. During inference, the model can retrieve semantically relevant information to assist in answering questions, thus compensating for its limitations in factual recall. Another method is a hybrid framework which integrates a privately fine-tuned SLM with a general-purpose LLM was proposed.

While SFT and RAG are widely used for adapting LLMs to domain-specific tasks, they face notable limitations in reasoning-intensive and privacy-sensitive settings. SFT tends to overfit surface forms, while RAG raises privacy concerns due to external query exposure. These issues are especially problematic for ToM tasks such as mental health consultations, where understanding latent user intent is critical. To this end, we propose a systematic comparison of existing adaptation strategies, and advocate for a hybrid SLM–LLM setup as a practical alternative. Our contributions include a comparative analysis of three paradigms, a focused discussion on ToM challenges, and a privacy-preserving design tailored for high-stakes dialog.

2 Related Work

Supervised Fine-Tuning As LLMs are increasingly adopted across a wide range of applications, the need for domain adaptation has become critical. SFT has emerged as a widely used approach to meet the demands of specialized domains [Yu *et al.*, 2025]. Researchers have observed that domain-specific tasks—such as medical diagnosis, legal consultation, or advanced mathematical reasoning—often require knowledge that goes beyond the general capabilities of pre-trained models [Huang *et al.*, 2023; Chen *et al.*, 2023]. In these scenarios, vanilla pre-trained LLMs may struggle to produce accurate or contextually appropriate responses. To address this gap, domain practitioners typically fine-tune pre-trained models on curated, task-relevant datasets, yielding domain-adapted models that better align with specific application requirements. This fine-tuning process enables the model to internalize domain-specific terminology, structures, and reasoning patterns, thereby improving both the relevance and quality of its outputs in specialized contexts.

Retrieval-Augmented Generation In addition to fine-tuning, RAG offers an alternative paradigm for domain adaptation that circumvents the high computational and resource costs associated with training large models. Rather than modifying the model parameters, RAG architectures attach an external knowledge base to an LLM, enabling the model to dynamically access relevant information during inference [Su *et al.*, 2025]. When faced with a user query, a retriever module [Shi *et al.*, 2023] is employed to identify and fetch semantically relevant documents from the knowledge base. These retrieved passages are then incorporated into the model's input, enriching the context available to the LLM. This allows even a general-purpose, non-domain-specific model to access and reason over domain-relevant knowledge on-the-fly. Extensive research has been devoted to improving the effectiveness of RAG systems. Efforts include enhancing the retrieval component—for example, using dense or hybrid retrievers to improve recall of relevant documents—as well as curating higher-quality knowledge corpora to ensure that the retrieved content contributes meaningfully to response accuracy. These improvements aim to bridge the gap between general models and domain-specific requirements without incurring fine-tuning overhead.

3 Limitations of Existing Approaches

Limitations of SFT in Privacy-sensitive Scenarios

To address the tension between privacy and performance in LLM-based applications, several strategies have been proposed—each offering partial solutions but also exhibiting critical limitations. As illustrated in Figure 1, 1Method (a), which relies solely on SLMs for both query processing and response generation, appears to offer strong privacy guarantees due to local deployment. However, even with SFT, SLMs inherently struggle with complex reasoning and open-ended tasks due to their constrained capacity and limited model size. On the other hand, 1Method (b), which depends entirely on cloud-based LLMs, can achieve stronger overall performance but inevitably requires transmitting sensitive user data to external servers, posing serious privacy and confidentiality risks. 1Method (c), deploying an LLM locally, theoretically addresses both privacy and performance, but in practice, the immense scale of state-of-the-art LLMs—such as Meta's 405B-parameter LLaMA-3.1 or GPT-4, estimated at approximately 1.8 trillion parameters [Deroy and Maity, 2025; Brown and others, 2023]—renders this approach infeasible for deployment on edge or resource-constrained environments. Their substantial memory footprint and computational demands typically necessitate cloud infrastructure, introducing additional concerns related to latency and operational costs [Kandala *et al.*, 2024; Qu *et al.*, 2025; Jang and Morabito, 2025]. Furthermore, there are implicit challenges associated with directly using models that have not undergone domain-specific SFT. These pre-trained models usually possess overly general knowledge, lacking the necessary context for domain-specific reasoning tasks, resulting in suboptimal downstream performance. Although SFT can mitigate this issue by adapting models to domain-specific tasks, fine-tuning large-scale LLMs remains prohibitively expensive and technically challenging, even when utilizing parameter-efficient fine-tuning (PEFT) techniques. The high computational resources, memory bandwidth, and energy required further exacerbate the impracticality of SFT-based approaches, particularly for institutions lacking advanced computational infrastructure.

Limitations of RAG in Privacy-sensitive Scenarios Another mainstream approach, RAG, also suffers from practical limitations. First, collecting a sufficiently comprehensive and high-quality knowledge base is often nontrivial. In specialized domains such as healthcare or

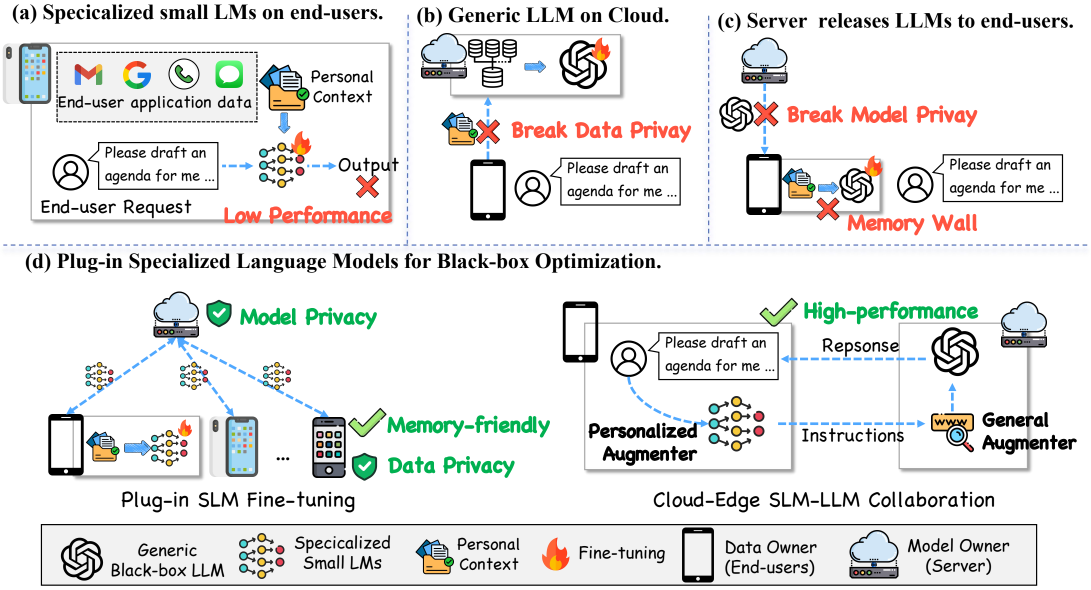
(a) Specicalized small LMs on end-users.
End-user application data
Personal Context
Please draft an agenda for me ...
Output
End-user Request
Low Performance
(b) Generic LLM on Cloud.
Break Data Privay
Please draft an agenda for me ...
(c) Server releases LLMs to end-users.
Break Model Privay
Please draft an agenda for me ...
Memory Wall
(d) Plug-in Specialized Language Models for Black-box Optimization.
Model Privacy
Memory-friendly
Data Privacy
Plug-in SLM Fine-tuning
High-performance
Please draft an agenda for me ...
Repsonse
Personalized Augmenter
Instructions
General Augmenter
Cloud-Edge SLM-LLM Collaboration
Generic Black-box LLM
Specicalized Small LMs
Personal Context
Fine-tuning
Data Owner (End-users)
Model Owner (Server)

Figure 1: Comparison of existing approaches. (a) Traditionally, specialized and smaller LMs (SLMs) on device are data privacy-friendly but underperformance. (b) End-users send personal context to the model owner (server) for fine-tuning. LLMs excel with context, but risk data privacy. (c) Server publishing LLMs to edge devices threatens proprietary model privacy, and it's not affordable for end users due to resource constraints. (d) Collaborating LLMs and SLMs enhances privacy and performance.

finance, data may be scarce, domain-specific, or sensitive—making it difficult to obtain or share. For applications involving sensitive data—such as in mental health, finance, or legal domains—reliance on external servers raises serious concerns regarding data exposure and regulatory compliance. These limitations highlight the growing mismatch between the capabilities of LLMs and the practical requirements of privacy-sensitive, real-world deployments. Second, the effectiveness of RAG heavily depends on the performance of the retriever module, which is responsible for accurately identifying and selecting relevant information from the database. Without a strong retriever, the model's overall performance degrades significantly. However, building or fine-tuning a high-quality retriever is itself a challenging task and often requires large amounts of annotated data and careful system design.

4 SLM–LLM Collaboration

SLM–LLM Collaboration Strikes a Balance Between Privacy and Reasoning.

Given these challenges, we argue that a hybrid SLM–LLM framework offers a more practical and privacy-preserving alternative. To reconcile the trade-off between privacy and performance, recent research has explored collaborative frameworks that combine SLMs and LLMs. In such designs, the SLM—owing to its compact size and ease of deployment on edge devices—serves as a lightweight, privacy-aware front-end. As illustrated in 1(d), the specialized SLMs preprocess or augments user queries using sensitive, domain-specific knowledge through supervised fine-tuning on local data. The query could also be further enhanced, by some local aurgumenters, which could helping enhance the quality of enhanced queries. Then the enriched query is passed to the LLM, which leverages its powerful generalization and reasoning capabilities to produce the final response.

This two-stage collaboration harnesses the complementary strengths of both models: the privacy protection and domain specialization provided by the locally deployed SLM, and the robust generative capability of the general-purpose LLM. Such a hybrid architecture effectively balances domain specificity, computational feasibility, and privacy preservation, aligning well with high-stakes applications requiring both robust reasoning and strict confidentiality.

5 Case Study: Mental Health

In mental health applications, understanding a user's latent mental state—such as emotional distress, intent, or unspoken fear—is essential [Chung *et al.*, 2023]. These are classic ToM tasks, where the system must infer beliefs and intentions from language that is often indirect, ambiguous, or context-dependent. While both SFT and RAG have shown promise in various domains, they present serious limitations in high-stakes mental health settings. Effective fine-tuning of an LLM via SFT necessitates access to large volumes of sensitive conversational data, such as therapy transcripts or patient records. However, acquiring and curating such data poses ethical and legal challenges due to confidentiality constraints and privacy regulations like HIPAA or GDPR [Sawhney *et al.*, 2022; Mandal *et al.*, 2025]. Even if such data were available, deploying SFT-tuned models would often require running large LLMs in centralized servers, which increases the risk of unintended data exposure.

RAG suffers from similar limitations [Zeng *et al.*, 2024]. Although it avoids fine-tuning the model directly, RAG requires user queries to be sent to external retrievers or knowledge stores, which still poses a significant privacy threat. In therapeutic scenarios, the contextual understanding often relies on a large amount of user-specific history or mental health background. Attaching such information to LLM input—whether directly or through retrieved context—effectively results in transmitting private data outside the user's device. This defeats the goal of privacy preservation. Moreover, RAG systems typically rely on pre-constructed document collections, which may lack appropriate coverage or nuance for mental health inference, further compounding hallucination risks and reducing the system's ability to reason about users' latent mental states.

By contrast, the hybrid SLM–LLM design keeps all privacy-sensitive processing local to the user's device. It enables mental state reasoning and contextualization to occur without leaking sensitive user inputs, while still benefiting from the LLM's general reasoning strength in a controlled, privacy-aware manner. This hybrid SLM–LLM framework offers a privacy-preserving alternative. The SLM, deployed on the user's device, is fine-tuned on annotated therapeutic dialoguesm. It is responsible for detecting emotional cues (e.g., avoidance, despair), disambiguating the user's intent, and generating a structured intermediate representation—such as a summarized psychological intent label or reformulated query. This enriched query is then forwarded to the cloud-hosted LLM, which uses its extensive generalization capabilities to generate empathetic, context-appropriate responses. For example, if the SLM identifies that a user is expressing avoidance regarding social situations, the LLM can incorporate cognitive-behavioral strategies into its response. This pipeline preserves user privacy by keeping sensitive data local and reduces hallucinations by grounding the LLM's response on SLM-inferred structure. It also allows ToM-style reasoning to be performed on-device, ensuring the system can interpret and respond to mental state cues responsibly.

6 Conclusion

In this work, we conduct a comprehensive comparison of existing approaches aimed at preserving user privacy and reducing hallucinations in LLMs, with a particular focus on their application in ToM tasks such as mental health consultation. Through both systematic analysis and empirical investigation, we examine the limitations of conventional solutions such as SFT and RAG, especially in privacy-sensitive, reasoning-intensive settings. Based on our findings, we argue that a hybrid framework combining a locally fine-tuned SLM with a powerful general-purpose LLM offers a more effective and privacy-compliant alternative. This collaborative setup not only mitigates hallucinations by enriching context with domain-specific knowledge, but also enables more personalized, context-aware responses—thereby better understanding user intent. Our results highlight the potential of SLM–LLM cooperation as a promising direction for building trustworthy and high-performance AI systems in sensitive human–AI interaction scenarios.

References

[Brown and others, 2023] Tom Brown et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2023.

[Chandra *et al.*, 2025] Mohit Chandra, Siddharth Sriraman, Harneet Singh Khanuja, Yiqiao Jin, and Munmun De Choudhury. Reasoning is not all you need: Examining llms for multi-turn mental health conversations, 2025.

[Chen *et al.*, 2023] Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*, 2023.

[Chu *et al.*, 2025] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025.

[Chung *et al.*, 2023] Neo Christopher Chung, George Dyer, and Lennart Brocki. Challenges of large language models for mental health counseling, 2023.

[Deroy and Maity, 2025] Aniket Deroy and Subhankar Maity. Code generation and algorithmic problem solving using llama 3.1 405b, 2025.

[Huang *et al.*, 2023] Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report, 2023.

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

[Jang and Morabito, 2025] SiYoung Jang and Roberto Morabito. Edge-first language model inference: Models, metrics, and tradeoffs. *arXiv preprint arXiv:2505.16508*, 2025.

[Kandala *et al.*, 2024] Savitha Viswanadh Kandala, Pramuka Medaranga, and Ambuj Varshney. Tinyllm: A framework for training and deploying language models at the edge computers, 2024.

[Kwan *et al.*, 2024] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024.

[Laban *et al.*, 2025] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

[Mandal *et al.*, 2025] Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. Towards privacy-aware mental health ai models: Advances, challenges, and opportunities, 2025.

[Mirzadeh *et al.*, 2024] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024.

[Qu *et al.*, 2025] Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. Mobile edge intelligence for large language models: A contemporary survey, 2025.

[Sawhney *et al.*, 2022] Ramit Sawhney, Atula Tejaswi Neerkaje, Ivan Habernal, and Lucie Flek. How much user context do we need? privacy by design in mental health nlp application, 2022.

[Shi *et al.*, 2023] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

[Su *et al.*, 2025] Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. *arXiv preprint arXiv:2501.15915*, 2025.

[Yao *et al.*, 2025] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination?, 2025.

[Yu *et al.*, 2025] Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training, 2025.

[Zeng *et al.*, 2024] Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag), 2024.