Poster: Towards Privacy-Preserving and Personalized Smart Homes via Tailored Small Language Models

Xinyu Huang, Leming Shen, Zijing Ma, Yuanqing Zheng
The Hong Kong Polytechnic University,, Hong Kong SAR, China
{unixy-xinyu.huang,leming.shen,zijing.ma}@connect.polyu.hk,csyqzheng@comp.polyu.edu.hk,

ABSTRACT

Large Language Models (LLMs) exhibit remarkable language comprehension to revolutionize smart homes. Existing LLM-based smart home assistants typically transmit user commands, along with user profiles and home configurations, to remote servers to obtain personalized services. However, users are increasingly concerned about potential privacy leakage. To address this, we develop *HomeLLaMA*, an on-device assistant for privacy-preserving personalized smart homes with a tailored small language model (SLM). *HomeLLaMA* learns from cloud LLMs to deliver satisfactory responses and enable user-friendly interactions. Once deployed, *HomeLLaMA* facilitates proactive interactions by continuously updating local SLMs and user profiles. To further enhance user interaction while protecting privacy, we develop *PrivShield* to offer an optional privacy-preserving serving for those users who are unsatisfied with local responses and willing to send less-sensitive queries to remote servers. Experiments demonstrate *HomeLLaMA* provides satisfactory services while significantly enhancing user privacy.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Security and privacy** → **Human and societal aspects of security and privacy**.

KEYWORDS

Smart Home, Large Language Model, Privacy-Preserving

ACM Reference Format:
Xinyu Huang, Leming Shen, Zijing Ma, Yuanqing Zheng . 2025. Poster: Towards Privacy-Preserving and Personalized Smart Homes via Tailored Small Language Models. In *The 31st Annual International Conference on Mobile Computing and Networking (ACM MOBICOM '25), November 4–8, 2025, Hong Kong, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3680207.3765677

ACM MOBICOM '25, November 4–8, 2025, Hong Kong, China

ACM ISBN 979-8-4007-1129-9/25/11…$15.00
https://doi.org/10.1145/3680207.3765677

Figure 1: The performance-privacy dilemma for users.

1 INTRODUCTION

The rapid advancement of smart homes has enabled intelligent and convenient living through AI and IoT technologies [1]. However, current commercial assistants such as Apple Siri [2, 3] rely heavily on predefined command-action mappings, limiting their adaptability to handle freestyle commands. Recent studies [4] have explored using large language models (LLMs) to better understand natural language commands and generate more flexible action plans. While effective, such cloud-hosted models expose users to potential privacy breaches under the honest-but-curious threat model. On the other hand, deploying small language models (SLMs) locally for serving smart homes preserves privacy but suffers from performance degradation due to the limited model capability. Therefore, as illustrated in Fig. 1, there exists a performance-privacy dilemma for smart home users.

To address this dilemma, we propose *HomeLLaMA*, a privacy-preserving local home assistant that delivers personalized services through continuous learning. The key insight of *HomeLLaMA* is empowering local SLMs with the capabilities of cloud LLMs to shift most privacy-sensitive query processing tasks from the cloud to the local. The powerful cloud services can be consulted with users' explicit approval only when needed. Specifically, *HomeLLaMA* features three key technical modules: *Local SLM Enhancement* for enhancing the performance of local assistants with a

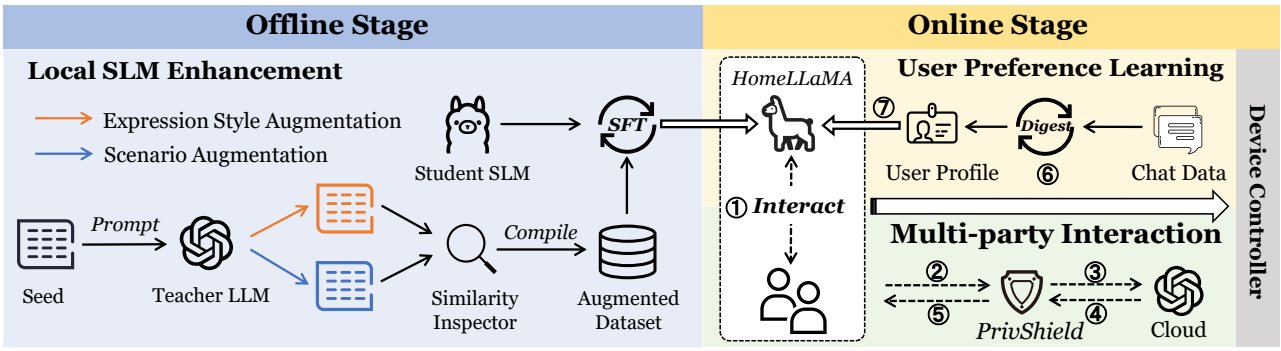


Figure 2: System overview of *HomeLLaMA*.

tailored inference paradigm, *Local-Cloud Collaboration* for maximizing user experience by seeking cloud LLMs for help in a privacy-preserving manner when users are unsatisfied with the current responses, and *User Preference Learning* for efficient locally-hosted long-term personalized services.

We implement *HomeLLaMA* on a local server and evaluate its performance. Extensive experiments show that *HomeLLaMA* enhances user privacy while maintaining acceptable performance, resolving the raised dilemma.

2 SYSTEM DESIGN

2.1 Local SLM Enhancement

HomeLLaMA first follows an offline stage to enhance the capability of local SLMs. We investigate the potential of leveraging powerful cloud LLMs to automatically synthesize a customized dataset based on a small amount of collected data (*i.e.*, command pool). This approach effectively transfers the knowledge embedded within the cloud LLM (teacher) to the local SLM (student) through the fine-tuning process.

Synthesis of commands. During each iteration of synthesis, we randomly sample five commands from the collected command pool as a starting point. We proceed to augment the original commands along the following two directions:

- *Vertical synthesis* generates new commands for different smart home scenarios. With the sampled seed commands, we instruct the cloud LLM to generate a new yet relevant command considering a different scenario.
- *Horizontal synthesis* aims to generate new commands with varied expression styles. Similar to the vertical synthesis process, we instruct the cloud LLM to change the expression style of the original command while maintaining its original meaning.

Similarity inspector. To ensure the diversity of augmented commands, we remove semantically redundant candidates by comparing each new command against existing ones based on the ROUGE-L similarity. A new command is retained only if its maximum similarity with any existing command falls below a predefined threshold α.

Inference paradigm. To serve diverse home environments, we introduce a tailored inference paradigm using Chain-of-Thoughts. The SLM is prompted to first generate a comprehensive set of relevant devices, assuming a fully equipped home with almost all COTS devices. Then, to adapt the results to a specific home, the system matches the generated device set with the actual available devices, producing a customized plan aligned with the user's environment.

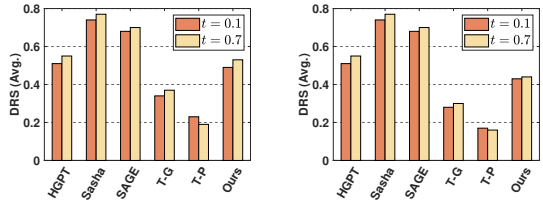
The tailored inference paradigm is first employed to label commands for fine-tuning the local SLM. After enhancement, the same paradigm is reused during inference to activate the model's improved capability in identifying relevant devices.

2.2 Multi-Party Interaction

User-assistant interaction. Upon receiving a command, the assistant generates an action plan and awaits user feedback. Users may *accept*, triggering immediate device control; provide *advice* in natural language to refine the plan; or *reject* it, prompting the assistant to invoke the cloud LLM for improved planning via the local-cloud collaboration module.

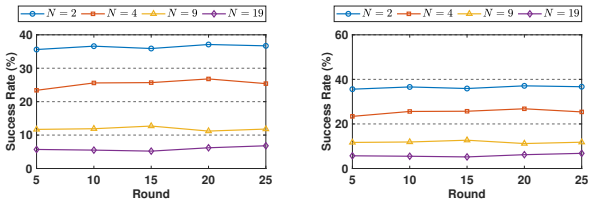
Local-cloud collaboration. When a user rejects a plan, *HomeLLaMA* may seek permission to consult a cloud LLM for improved response generation. Upon approval, the assistant initiates the collaboration while considering user privacy. In practice, all user profiles and smart home configurations remain local, and only the current command is transmitted externally. To further minimize privacy risks, *HomeLLaMA* employs *PrivShield*, a lightweight obfuscation module that anonymizes the user query before it is sent to the cloud.

PrivShield. Operating in a **SLM-in-the-middle** framework, *PrivShield* consists of three privacy-preserving steps: (1) rewriting user commands to remove specific personal details and redundancy using the local SLM [5], (2) generating multiple unrelated adversarial commands to obfuscate the original user command, and (3) recovering the response corresponding to the original user command after querying the cloud. The final cloud-enhanced plan is then adapted locally by the SLM and presented to the user as the improved plan.



(a) FP16 precision of SLMs. (b) INT8 precision of SLMs.

Figure 3: *DRS* results in (a) FP16 and (b) INT8 precision.



(a) Phi3-3B. (b) LLaMA3-8B.

Figure 4: ASR results of two LLMs across query rounds.

2.3 User Preference Learning

Profile generation. *HomeLLaMA* records interaction histories locally and summarizes each conversation into a concise user profile. Each profile contains key topics, preferences, user commands, and the final approved plans, and is stored in a text embedding database.

Profile updating. To prevent redundancy, newly generated profiles are compared with existing ones using cosine similarity. If sufficiently distinct, they are stored as new entries; otherwise, they are merged into existing profiles through guided SLM prompts, enhancing storage efficiency.

Personalized plan generation. During inference, the assistant retrieves the top-matching profiles based on similarity to the new query. These are decoded and combined with the current command and home configuration to produce the final personalized action plans for users.

3 EVALUATION

Implementation. *HomeLLaMA* is deployed on a local server. We use GPT-4-Turbo to augment 14K command-action pairs from IFTTT seeds across 9 smart home scenarios, filtered with a ROUGE-L threshold of 0.7. Meta-LLaMA3-8B is fine-tuned via QLoRA for 3 epochs on an RTX 4090 GPU.

Quality of service. To evaluate *HomeLLaMA* and baselines, we propose a *DevFinder* benchmark covering various commands and scenarios. As shown in Fig. 3, while cloud-based assistants still achieve the highest overall scores, *HomeLLaMA* delivers comparable device relevance scores (DRS) to GPT-3.5, demonstrating strong performance with strict local privacy protection via *PrivShield*. It also outperforms other on-device models such as TT-Gemma and TT-Phi-2, benefiting from the domain-specific fine-tuning. Additionally, it can be concluded that increasing the model temperature can enhance performance but may also introduce hallucinations, with smaller models experiencing performance degradation due to their limited reasoning capabilities.

Privacy protection. To evaluate privacy under the honest-but-curious model, we launch activity monitoring attacks by prompting GPT-4 to infer user activities from inputs. We use the *DevFinder* dataset and report the *attack success rate (ASR)* as the metric. We vary the number of adversarial commands and the strength of base SLMs to analyze their effects. As shown in Fig. 4, *PrivShield* consistently lowers ASR across all settings, validating its effectiveness in anonymizing cloud queries. Injecting more adversarial commands introduces semantic noise that confuses the attacker, thereby improving protection. Stronger SLMs further enhance this effect by generating more diverse and obfuscated outputs. However, both strategies may increase response latency and computational overhead, suggesting the importance of tunable privacy-performance trade-offs.

4 CONCLUSION

This paper presents an on-device smart home assistant that balances privacy and performance for users. The designed system comprises three modules to develop a novel local-cloud integration, enabling seamless and privacy-enhanced user interactions. Extensive experiments verify the effectiveness of *HomeLLaMA* in enhancing user privacy while keeping acceptable performance in the quality of services.

ACKNOWLEDGMENTS

We sincerely thank our anonymous reviewers for their constructive comments and invaluable suggestions that helped improve this paper. This paper is supported by Hong Kong GRF under Grant No. 15206123 and 15211924. Yuanqing Zheng is the corresponding author.

REFERENCES

[1] H. Xu, L. Han, Q. Yang, M. Li, and M. Srivastava, "Penetrative ai: Making llms comprehend the physical world," in *ACM HotMobile 2024*, pp. 1–7, 2024.

[2] A. S. Tulshan and S. N. Dhage, "Survey on virtual assistant: Google assistant, siri, cortana, alexa," in *Advances in Signal Processing and Intelligent Recognition Systems*, pp. 190–201, Springer, 2019.

[3] X. Huang, L. Shen, Z. Ma, and Y. Zheng, "Towards privacy-preserving and personalized smart homes via tailored small language models," *arXiv preprint arXiv:2507.08878*, 2025.

[4] E. King, H. Yu, S. Lee, and C. Julien, "Sasha: creative goal-oriented reasoning in smart homes with large language models," *ACM IMWUT*, vol. 8, no. 1, pp. 1–38, 2024.

[5] L. Shen, Q. Yang, X. Huang, Z. Ma, and Y. Zheng, "Gpiot: Tailoring small language models for iot program synthesis and development," in *ACM SenSys*, pp. 199–212, 2025.