Guardians of Privacy: Leveraging LLMs in Assistive Robotic Systems for Healthcare

ZOUGHALIAN, Kavyan, AITSAM, Muhammad, MARCHANG, Jims <http://orcid.org/0000-0002-3700-6671> and DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>

Guardians of Privacy: Leveraging LLMs in Assistive Robotic Systems for Healthcare

Kavyan Zoughalian, Muhammad Aitsam, Jims Marchang, Alessandro Di Nuovo
Department of Computing, Sheffield Hallam University
Advanced Wellbeing Research Centre (AWRC), Sheffield, United Kingdom
Emails: {k.zoughalian, m.aitsam, j.marchang, a.dinuovo}@shu.ac.uk

***Abstract*—This paper presents a Privacy-Aware Assistive System (PAAS) designed to mitigate privacy risks associated with integrating cloud-based Large Language Models (LLMs) into resource-constrained systems like Socially Assistive Robots (SARs), in healthcare settings, which results in seeking cloud-based solutions. While cloud-based LLMs substantially enhance patient care through more natural and effective interactions, their use raises significant privacy concerns due to the potential exposure of sensitive healthcare data to external services. To address this challenge, PAAS employs the Principle of Least Privilege (PoLP), by fine-tuning domain-specific LLMs to accurately identify user intents and extract necessary parameters used to query a structured database from unstructured natural language inputs without exposing sensitive data directly to third-party services. The paper introduces an algorithm for generating domain-specific datasets, facilitating precise intent classification and custom entity recognition essential for querying internal databases securely. Performance comparisons among fine-tuned models were conducted using varying complexities of user requests, including basic, context-sensitive, and ambiguous interactions. Results demonstrate robust performance, with the GPT-4O-mini fine-tuned model achieving an F1 score of up to 95% across multiple tests conducted at different times and days. PAAS effectively facilitates high-quality, natural user interactions through the advanced capabilities of LLMs while rigorously maintaining user privacy. Future research will address improving the system's resilience to linguistic ambiguities and further advancing its privacy safeguards.**

***Index Terms*—large language models, socially assistive robots, privacy-aware systems, Artificial intelligence in healthcare**

I. Introduction

Assistive robots already reduce falls, medication errors, and caregiver workload in elder-care and rehabilitation settings [1], [2]. Socially assistive robots (SARs) amplify this benefit by holding natural conversations that improve user engagement and adherence [3], [4]. Recent large language models (LLMs) such as ChatGPT and LLaMA bring state-of-the-art dialogue capacity to SARs, enabling context-aware, personalised coaching and triage [5]–[8].

A substantial share of clinical data, laboratory values, medication orders, and progress notes reside in relational electronic health record (EHR) warehouses queried through SQL. Translating a plain request (e.g., "*show potassium >5 mmol/L for the past 24 h*") into SQL allows non-technical staff to access insights directly [9]. However, state-of-the-art LLMs contain billions of parameters and exceed the compute envelope of the on-board SAR hardware; Cloud LLM-as-a-Service (LLMaaS) offers a feasible alternative if privacy and sub-2 s latency can be guaranteed [10].

We therefore propose the **Privacy-Aware Assistive System (PAAS)**. PAAS fine-tunes a base LLM on healthcare intents, enforces *least-privilege* data exchange so only anonymised query fragments leave the robot, and returns structured output that is used to execute locally against the EHR. This architecture balances conversational flexibility with regulatory requirements such as HIPAA and GDPR, validated on three increasingly complex clinical query sets.

The main contributions of this paper include the following:

- A novel algorithm to generate domain specific training dataset for fine-tuning LLMs to process healthcare-related, unstructured user queries. Although prior work has explored LLM wrappers for domain-specific datasets, our system couples least-privilege data redaction and robot-side custom intent extraction, used for on-prem SQL execution.
- A PAAS that adheres to the principle of least privilege when interacting with cloud-based LLMs in healthcare settings, only leveraging LLM to retrieve what is required.
- A performance analysis comparing the proposed PAAS with standard LLM models, demonstrating its ability to preserve privacy without sacrificing performance.

PAAS securely mediates SAR interactions with cloud LLMs. It auto-generates task-specific data, fine-tunes the models, and enforces the Principle of Least Privilege so requests are parsed accurately without exposing sensitive information. As Figure 1 shows, this privacy-first layer is key to integrating SARs with healthcare databases and cloud-based LLMs while safeguarding patient data.

The implementation code, the dataset used to train and test the models, are publicly available at our GitHub repository [11] and the paper's dedicated webpage [12].

Section II surveys prior work; III reviews LLM privacy risks; IV details our fine-tuning algorithm; V introduces PAAS and test datasets; VI reports performance vs. baseline LLMs; VII concludes.

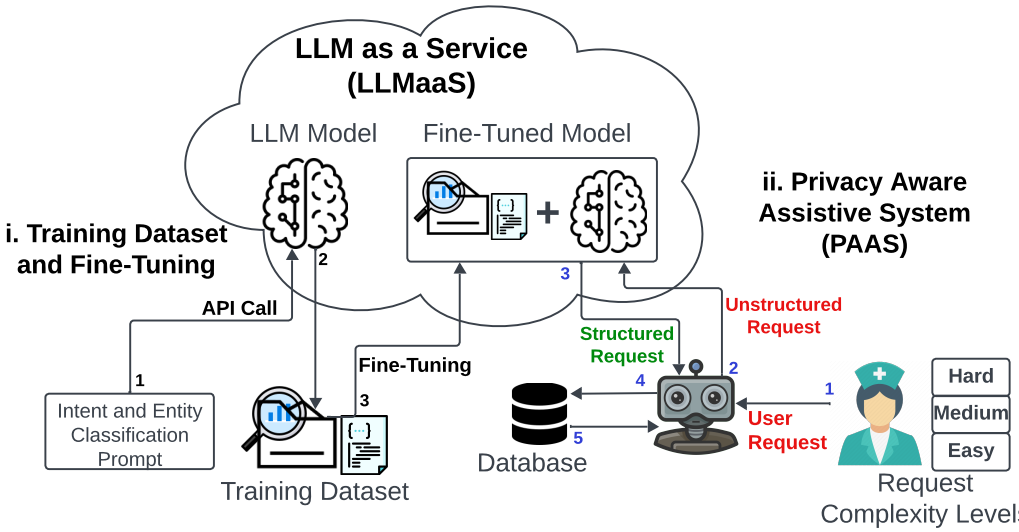

Figure 1: The figure illustrates the overall process involved in creating the PAAS. The generation of the training dataset for fine-tuning is shown on the left side of the figure, while the right side depicts how the fine-tuned model is utilized by PAAS for natural language understanding of user requests, adhering to the principle of least privilege.

II. Related Work

A. Natural-Language Access to Clinical Databases

Early systems relied on hand-crafted templates, for example, the classic railway enquiry demo, which achieved near-deterministic accuracy yet failed on linguistic variability and domain shift. Modern solutions fall into three families.

Neural semantic parsing: RNN and transformer parsers map utterances directly to SQL. MedT5SQL, a T5-based model fine-tuned on radiology reports, pushes accuracy to 88 % on its in-domain test set but requires tens of thousands of labelled pairs and lacks privacy safeguards [13]. Hallucinations remain a risk [14].

Retrieval-augmented generation (RAG): RAG pipelines mitigate hallucination by injecting schema documentation at inference time, yet incur a heavy GPU costs on-device or exposes PHI when run in the cloud [15]. Recent systems papers report that retrieval accounts for roughly 40 % of end-to-end latency, doubling the time-to-first-token (TTFT) in a 10-million-chunk FAISS pipeline [16]

Large language models: Seq2seq LLMs—T5, GPT-3.5, GPT-4-now top the Spider leaderboard, achieving >85 % exact-match with few-shot prompting [17]. Sung et al. extend this idea by auto-generating NL-SQL pairs via GPT-4 [18], but their pipeline omits (i) privacy as they provide highly sensitive documents to the LLMs, (ii) lack of task-complexity labels.

B. Modular NLU for Safety-Critical Domains

Modular stacks decompose a query into intent, entities, and domain class [19]. Although less flexible than end-to-end LLMs, they offer auditability and can refuse incomplete requests, a property our Privacy-Aware Assistive System (PAAS) retains by routing ambiguous queries to a clarification workflow.

III. Privacy Issues in LLMs

Privacy concerns regarding interactions with LLMs, such as GPT-based systems, remain underexplored despite their growing significance in artificial intelligence (AI). While technical capabilities and ethical issues of LLMs have been extensively studied, privacy considerations often receive limited attention. LLMs, trained on vast datasets containing personal information, pose risks related to data retention, potential breaches, and misuse [6].

Recent studies have highlighted privacy challenges, demonstrating that LLMs can inadvertently expose sensitive or personally identifiable information from their training data [20]–[22]. This indicates that data leakage is a genuine concern, despite limited acknowledgment. Approaches such as selective privacy-preserving fine-tuning [23] and task-specific knowledge distillation combined with differential privacy [24] have been proposed to mitigate these risks. However, the scalability and effectiveness of such methods in large-scale deployments require further investigation and remain actively debated [25].

LLM providers have implemented policies to address privacy concerns; however, these policies often lack clarity and transparency. For instance, OpenAI anonymizes user data and requires explicit consent for model training, yet retains conversations for 30 days, posing potential security risks [26], [27]. OpenAI's unclear policies regarding data retention across model updates, coupled with ChatGPT's memory feature storing user-specific data, further complicate privacy concerns [28]. Similarly, Google's Bard retains user data for at least 18 months, highlighting inconsistencies among providers [29]. Existing measures primarily address individual interactions but inadequately protect against advanced threats like data re-identification. Thus, further research into transparent and scalable privacy frameworks for LLM technologies is critical.

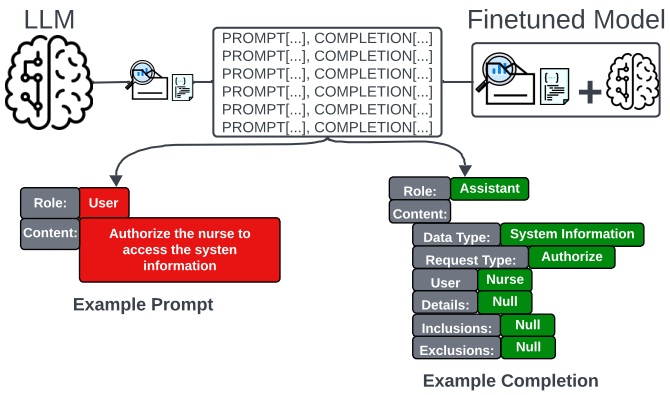

Figure 2: Fine-tuning training dataset generated using the algorithm, consisting of prompt-completion pairs.

IV. Fine-tuning LLMs

Fine-tuning LLMs involves adapting a model to specific tasks or datasets by updating its parameters with additional training data. This process leverages existing knowledge from initial training on large datasets, significantly improving performance in specialized natural language processing (NLP) tasks such as named entity recognition (NER), sentiment analysis, and question-answering [30]. Particularly useful when labelled data is limited, fine-tuning allows models to combine prior knowledge with new information, enhancing their ability to identify and classify domain-specific entities accurately [31], [32]. Thus, the quality and relevance of training data used in fine-tuning are critical to achieving optimal performance.

Algorithm 1 Creating the Fine-Tuning GPT Model training-dataset for Socially Assistive Robot

Load required libraries
Initialize OpenAI client with the API key
Define prompt template for the SAR:
    The model responds in JSON format
    The response includes the following entity classes:
    `x={\entity1: description, entity2...}`
    `y={\entity3: description, entity4...}`
    Entities that are not provided but required can be `Null`
procedure ProcessUserRequest(input_request)
    Generate a prompt based on the defined template
    Call GPT model via API to generate a response
    Return response in JSON format
end procedure
for each user request in the dataset **do**
    Call `ProcessUserRequest` with the current request
    for system, user, assistant **do**
        Define `content`
    end for
    Log the response and save it to the training dataset file
end for

As highlighted in the literature, a significant bottleneck when fine-tuning large language models (LLMs) is curating a domain-specific training set, an exercise that is typically costly and time-consuming. In this *proof-of-concept study* we therefore limit ourselves to a deliberately small corpus: Algorithm 1 shows how we generate a 100-prompt training set and three 50-prompt test sets (one per complexity tier) with GPT-4. Figure 2 illustrates a typical prompt–completion pair. As our corpus is deliberately small, withholding a portion for validation would reduce effective training data by up to 30 %. We therefore follow OpenAI's guidance and train on the full set, using the built-in loss curve as a convergence check The core of the procedure is a reusable prompt template that instructs the LLM to emit JSON-formatted answers, enabling the socially assistive robot (SAR) to extract entities and act on them in a context-aware manner.

Entity classifications within the template are defined using a dictionary format, clearly instructing the LLM on expected entities from user requests. For instance, an entity classification might be 'Request_Type', with possible entities such as 'Data_Access' along with a descriptive context. Given the dynamic nature of user requests, the template includes rules allowing entities not explicitly mentioned to be assigned a 'Null' value, preventing the model from hallucinating or forcefully and inaccurately generating nonexistent entities. Consequently, entities not specified in the user query are explicitly set to 'Null' rather than invented.

The fine-tuning of models is performed using OpenAI's cloud-based Large Language Model as a Service (LLMaaS) [37]. In the fine-tuning process, the temperature parameter controls the creativity or randomness of the text generated. We set `temperature`=0.3 and `top_p`=0.95, following the "low-temperature, high-coverage" recipe recommended by OpenAI for deterministic paraphrase generation [38]. Higher values (0.5–0.7) introduced clinically implausible tokens; lower values (<0.2) produced near duplicates. The setting is therefore data-driven rather than arbitrary.

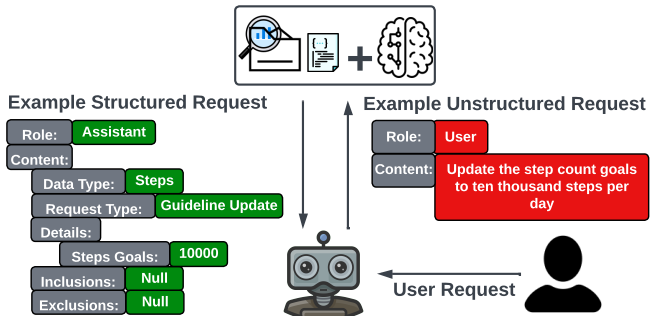

Figure 3: From an Unstructured User Request to Stuctured Request Using the Fine-Tuned LLM

A. Dataset Stratification by Complexity

We evaluate model robustness across three *complexity tiers* illustrated in Table I: *Easy* covers basic entity requests, *Medium* introduces overlapping or nested entities *Hard* focuses

Difficulty	Description	Examples
Easy	Basic Entity Recognition with straightforward user requests dealing with well-defined entities based on the works of [33] and [34].	Example 1: "Please provide access to the patient's heart rate data." Example 2: "I need to access the steps data of the patient."
Medium	Nested and Context-Sensitive user requests, overlapping or nested entities. Testing the model's ability to handle advanced structure and context-dependent decisions [35].	Example 1: "As a carer, I want to access the total sum of Steps data for the patient between 2024-05-10 and 2024-06-10, excluding yesterday." Example 2: "Please provide the average of the patient's heart rate data between 2024-07-20 and 2024-08-25, excluding data from 2024-08-01 to 2024-08-10."
Hard	Fine-Grained Entity Recognition, specific entities with some user request ambiguity [36].	Example 1: "Please update the threshold heart rate for the patient to five thousand as requested by the doctor." Example 2: "The nurse needs to modify the steps goals for the patient."

Table I: Entity Recognition Complexity Levels separated to basic entity recognition (easy) and nested context-sensitive entity recognition (medium) fine-grained entity recognition with ambiguity (hard)

on fine-grained entities with additional user-query ambiguity. Although the last tier is conceptually "harder," its difficulty lies in semantic ambiguity, an aspect not fully captured by token counts or brace depth. As shown in Table II, to provide a quantitative lens we measure:

1) **Prompt length** (ℓ): total input tokens, and
2) **Entity-nesting depth** (d): maximum brace depth in the ground-truth JSON.

A one-way ANOVA shows that both ℓ and d vary significantly across the three tiers (prompt length: $F2,147 = 685.24$, $p = 3.06\times10^{-75}$; nesting depth: $F2,147 = 2451.00$, $p = 1.30\times10^{-113}$), validating the stratification from a structural standpoint.

Table II: Descriptive statistics for the 150-query dataset (ℓ = prompt length, d = nesting depth).

Tier	#Queries	ℓ (mean $\pm$ SD)	d (mean $\pm$ SD)
Easy	50	8.7 $\pm$ 1.0	1.0 $\pm$ 0.0
Medium	50	20.4 $\pm$ 1.1	2.0 $\pm$ 0.0
Hard	50	13.8 $\pm$ 2.3	2.0 $\pm$ 0.1

Hard queries are not longer than Medium ones because their challenge is linguistic ambiguity rather than sheer length or structural depth; the quantitative metrics therefore understate their true difficulty. Conversely, medium queries stress the nested structure, and thus register a higher average prompt length. We treat the two upper tiers as *orthogonal* difficulty axes, lexical breadth versus contextual ambiguity, rather than a simple hierarchy.

V. Privacy-Aware Assistive System (PAAS)

The Principle of Least Privilege (PoLP) [39] mandates that users, programs, or processes should only be granted the minimum access necessary to perform their required functions. The Privacy-Aware Assistive System (PAAS) integrates PoLP by minimizing LLM access to sensitive information about the user; held in the database, enhancing user privacy, and adhering to ethical data handling practices, particularly crucial in healthcare, where protecting patient data is vital [40], [41].

Additionally, PAAS aligns with broader cybersecurity frameworks such as zero-trust architecture and privacy by design, emphasizing data minimization, continuous verification of user identities, and access rights. Zero-trust assumes threats exist both externally and internally, necessitating stringent access control and consistent enforcement of PoLP across organizational levels [42]. This approach strengthens security measures and helps ensure compliance with stringent data protection regulations, including the General Data Protection Regulation (GDPR).

As depicted in Figure 3, the Large Language Model (LLM) leverages Natural Language Processing (NLP) to accurately extract relevant entities within defined system capabilities. The structured outputs from the LLM enable efficient retrieval of necessary information to address user requests.

The system specifically includes six entity classifications tailored for socially assistive robots in healthcare, facilitating data access requests, updating well-being guidelines, and authorizing new user access. Performance metrics were evaluated at four different times and days, these are specified in the supplementary material on the repository. This approach ensures the robustness and reliability of results by accounting for variations due to network congestion and external factors. The network performance parameters are presented in Table III.

Table III: Network and ChatGPT API Performance Parameters

Network Parameters	Value
Data Rate (Upload Estimate Avg)	86 Mbps
Data Rate (Download Estimate Avg)	158 Mbps
Latency (Ping to API Avg)	18 ms
Traceroute Hops	7 hops

VI. Performance Analysis of privacy-preserving fine-tune LLM models

A. Performance Metrics and Error Analysis

We report five metrics *precision*, *recall*, F_1, token usage, and latency. Precision captures the reliability of positive predictions, while recall quantifies the detector's sensitivity. F_1 is their harmonic mean and serves as the primary ranking criterion.

A prediction is counted as a true positive (TP) when the predicted entity and its value match the ground truth exactly.

Table IV: Model Performance Across Dataset–Complexity Levels. Best values are in **bold**.

Metric	GPT-3.5-turbo		GPT-4o-mini	
	Fine-tuned	Base	Fine-tuned	Base
Easy				
Precision (%)	77.5	9.2	**95.0**	13.2
Recall (%)	91.2	16.0	**95.0**	18.0
F1 (%)	83.6	11.1	**95.0**	14.8
Prompt tokens	96	96	94	94
Completion tokens	57	76	**45**	128
Total tokens	153	172	**139**	222
Latency (s)	**1.94**	1.33	2.63	2.03
Medium				
Precision (%)	70.4	11.5	**70.6**	10.2
Recall (%)	**64.0**	10.9	63.7	8.0
F1 (%)	67.1	11.2	**67.0**	8.9
Prompt tokens	126	126	124	124
Completion tokens	87	124	**88**	158
Total tokens	213	250	**212**	282
Latency (s)	**1.51**	2.03	2.87	2.80
Hard				
Precision (%)	**59.2**	2.4	58.0	2.6
Recall (%)	**56.3**	3.7	56.2	3.4
F1 (%)	**57.8**	2.9	57.0	2.9
Prompt tokens	104	104	103	103
Completion tokens	57	97	**55**	129
Total tokens	161	201	**158**	232
Latency (s)	**1.27**	1.56	2.38	2.20

A true negative (TN) arises when both prediction and ground truth omit the entity. A false positive (FP) is recorded when the model asserts an entity that is absent or mismatched in the ground truth, and a false negative (FN) when the ground truth contains an entity the model misses.

Table IV contrasts the fine-tuned and base versions of GPT-3.5-turbo and GPT-4o-mini in three levels of task complexity. From the Easy subset, the fine-tuned GPT-4o-mini achieves 95% precision, recall, and F1 score, an across-the-board improvement of more than 80% over its baseline and at least 12% over the fine-tuned GPT-3.5-turbo. As complexity rises to the Medium subset, all models lose accuracy, but the fine-tuned variants remain ahead of their baselines. GPT-4o-mini beats GPT-3.5-turbo in precision and F1, although GPT-3.5-turbo posts the lowest latency (1.51 s). The Hard subset emphasizes the trend; both fine-tuned models drop to roughly 58% F1, yet still outperform their baselines by more than 50 points. Here, GPT-3.5-turbo preserves its latency advantage, while GPT-4o-mini maintains the smallest token footprint. Across the three subsets, the fine-tuned models consistently emit fewer completion tokens, often by 35% or more, than their respective baselines, reducing API cost and bandwidth. In aggregate, the fine-tuned GPT-4o-mini strikes the best balance of accuracy, efficiency, and latency, making it a strong candidate for real-time SAR deployments where both precision and resource usage are critical. These results underscore the value of fine-tuning for unstructured, domain-specific language understanding, while also highlighting the persistent challenge of ambiguity at higher task complexities.

B. Error Distribution and Metric Insights

The confusion matrices available on the website [12] visualises the raw error patterns that underlie the summary metrics in Table IV. Several observations emerge.

Easy dataset.: The resulting Precision and Recall converge at the upper end of the scale, which in turn drives the F_1 score toward its theoretical maximum of 1. Residual errors are dominated by *false positives*; manual inspection shows they arise when the model mistakenly labels templated greetings (e.g. "*Hello Doctor*") as clinically relevant entities. This error type explains why Precision (95 %) lags Recall by two percentage points for GPT-4o-mini.

Medium dataset.: TPs drop by 15–20 %, and both FP and FN increase relative to the Easy split. FP growth outpaces FN, so Precision degrades faster than Recall (GPT-3.5-turb$_{\text{ft}}$: 70.4 % vs. 64.0 %). Qualitative inspection reveals two root causes: (i) lexically ambiguous synonyms (*"tablet"* vs. *"pill"*) and (ii) nested entities in multi-clause queries. Because the harmonic mean penalises divergence between the two axes, F_1 settles in the high-60 range for the fine-tuned runs, still higher than the baseline models, but signalling that further disambiguation is needed.

Hard dataset.: FP and FN form $\sim$35 % of all predictions. False negatives increase more than false positives, so Recall erodes faster than Precision (GPT-4o-mini$_{\text{ft}}$: 56.2 % Recall vs. 58.0 % Precision). The confusions stem from deeply nested references and cross-sentence co-reference; e.g. resolving *"it"* to the correct medication when multiple candidates exist. Although each fine-tuned model still retains a higher F_1 margin over its baseline, the drop from Easy to Hard highlights the limits of zero-shot reasoning even after domain adaptation.

The matrices confirm that fine-tuning mostly reduces FP, hence boosts Precision on straightforward queries, while Recall suffers on ambiguity-heavy questions as FN rise. In a social-assistive-robot workflow, these asymmetric costs matter. An FP can annoy a user but a missed FN can jeopardise safety. Future work will therefore combine fine-tuning with lightweight retrieval-augmented generation to curb the FN spike observed at higher complexity levels.

VII. Conclusion

We introduced PAAS, a privacy-aware pipeline that applies the Principle of Least Privilege to large-language-model deployment in socially assistive robots. A lightweight algorithm generates 100 domain-specific prompts for fine-tuning, enabling the LLM to recognise healthcare entities and intents without exposing extra patient data. On a 150-query test suite, the fine-tuned GPT-4o-mini outperformed its baseline (and GPT-3.5) on precision, recall, and F1, while emitting fewer tokens, an attractive trade-off for resource-constrained robots. Performance falls on ambiguity-heavy queries, but both models remain usable across all tiers. These findings show that small, targeted fine-tuning can make cloud LLMs both accurate and privacy-respectful in clinical HRI. Next steps are to enlarge the synthetic dataset, add real clinical logs, and explore retrieval-augmented or local LLM variants to cope with deeper linguistic ambiguity.

Acknowledgment

For open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

[1] F. Cavallo, R. Esposito, R. Limosani, A. Manzi, R. Bevilacqua, E. Felici, A. Di Nuovo, A. Cangelosi, F. Lattanzio, P. Dario, *et al.*, "Robotic services acceptance in smart environments with older adults: user satisfaction and acceptability study," *Journal of medical Internet research*, vol. 20, no. 9, p. e9460, 2018.

[2] A. Di Nuovo, F. Broz, N. Wang, T. Belpaeme, A. Cangelosi, R. Jones, R. Esposito, F. Cavallo, and P. Dario, "The multi-modal interface of robot-era multi-robot services tailored for the elderly," *Intelligent Service Robotics*, vol. 11, pp. 109–126, 2018.

[3] L. Wang, Z. Wan, C. Ni, Q. Song, Y. Li, E. W. Clayton, B. A. Malin, and Z. Yin, "A systematic review of chatgpt and other conversational large language models in healthcare," *medRxiv*, 2024.

[4] J. Marchang, A. Di Nuovo, C. Elliott, H. Meese, S. Vinanzi, and M. Zecca, *Security and privacy in assistive robotics: cybersecurity challenges for healthcare*. EPSRC UK-RAS Network, 2023.

[5] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang, J. Gao, Y. G. S. Wang, J. ming Ji, Z. Qiu, M. Li, C. Qian, T. Guo, S. Ma, Z. Wang, Z. Guo, Y. Lei, C. Shao, W. Wang, H. Fan, and Y. D. Tang, "The application of large language models in medicine: A scoping review," *iScience*, vol. 27, p. 109713, 5 2024.

[6] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao, "Large language models for human–robot interaction: A review," *Biomimetic Intelligence and Robotics*, vol. 3, p. 100131, 12 2023.

[7] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging generative ai and large language models: A comprehensive roadmap for healthcare integration," *Healthcare (Switzerland)*, vol. 11, 10 2023.

[8] U. Mumtaz, A. Ahmed, and S. Mumtaz, "Llms-healthcare: Current applications and challenges of large language models in various medical specialties," *Artificial Intelligence in Health*, vol. 1, no. 2, pp. 16–28, 2024.

[9] X. Liu, S. Shen, B. Li, P. Ma, R. Jiang, Y. Zhang, J. Fan, G. Li, N. Tang, and Y. Luo, "A survey of nl2sql with large language models: Where are we, and where are we going?," *arXiv preprint arXiv:2408.05109*, 2024.

[10] Z. Cai, R. Ma, Y. Fu, and W. Zhang, "Llmaas: Serving large language models on trusted serverless computing platforms," *IEEE Transactions on Artificial Intelligence*, vol. PP, no. 99, pp. 1–11, 2024.

[11] K. Zoughalian, "Privacy-aware-system." https://github.com/kav-ros/Privacy-Aware-System, 2025. Accessed: 2025-06-07.

[12] K. Zoughalian, "Guardians of privacy: Leveraging llms in assistive robotic systems for healthcare." https://sites.google.com/view/kzcps2025/home. Accessed: 2025-06-07.

[13] A. Marshan, A. N. Almutairi, A. Ioannou, D. Bell, A. Monaghan, and M. Arzoky, "Medt5sql: a transformers-based large language model for text-to-sql conversion in the healthcare domain," *Frontiers in Big Data*, vol. 7, p. 1371680, 2024.

[14] S. Soni and K. Roberts, "Toward a neural semantic parsing system for ehr question answering," in *AMIA Annual Symposium Proceedings*, vol. 2022, p. 1002, 2023.

[15] R. Elgedawy, I. Danciu, M. Mahbub, and S. Srinivasan, "Dynamic q&a of clinical documents with large language models," *arXiv preprint arXiv:2401.10733*, 2024.

[16] M. Shen, M. Umar, K. Maeng, G. E. Suh, and U. Gupta, "Towards understanding systems trade-offs in retrieval-augmented generation model inference," *arXiv preprint*, vol. arXiv:2412.11854, Dec. 2024. Version 1, submitted 16 Dec 2024.

[17] D. Gao, H. Wang, Y. Li, X. Sun, Y. Qian, B. Ding, and J. Zhou, "Text-to-sql empowered by large language models: A benchmark evaluation," *arXiv preprint arXiv:2308.15363*, 2023.

[18] C.-W. Sung, Y.-K. Lee, and Y.-T. Tsai, "A new pipeline for generating instruction dataset via RAG and self fine-tuning," *48th IEEE Computers, Software and Applications Conference (COMPSAC 2024)*, 2024.

[19] H. Bhathena, A. Joshi, and P. Singh, "An efficient method for natural language querying on structured data," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 322–331, 2023.

[20] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, p. 100211, 6 2024.

[21] Q. Li, J. Hong, C. Xie, J. Tan, R. Xin, J. Hou, X. Yin, Z. Wang, D. Hendrycks, Z. Wang, B. Li, B. He, and D. Song, "Llm-pbe: Assessing data privacy in large language models," *Proceedings of the VLDB Endowment*, vol. 17, pp. 3201–3214, 7 2024.

[22] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

[23] T. Wang, L. Zhai, T. Yang, Z. Luo, and S. Liu, "Selective privacy-preserving framework for large language models fine-tuning," *Information Sciences*, vol. 678, p. 121000, 9 2024.

[24] S. Garg and V. Torra, "Task-specific knowledge distillation with differential privacy in llms," *Springer, Cham*, pp. 374–389, 2024.

[25] R. Behnia, M. Ebrahimi, J. Pacheco, and B. Padmanabhan, "Privately fine-tuning large language models with differential privacy," *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2022-November, pp. 560–566, 10 2022.

[26] OpenAI, "Openai privacy policy." https://openai.com/policies/privacy-policy/, 2023. Accessed: 2024-09-09.

[27] ChatGPT, "How chat retention works in chatgpt | openai help center."

[28] P. Silva, C. Gonçalves, N. Antunes, M. Curado, and B. Walek, "Privacy risk assessment and privacy-preserving data monitoring," *Expert Systems with Applications*, vol. 200, p. 116867, 8 2022.

[29] G. Bard, "Google bard - full privacy report."

[30] U. Katz, M. Vetzler, A. D. Cohen, and Y. Goldberg, "Neretrieve: Dataset for next generation named entity recognition and retrieval," *arXiv preprint arXiv:2310.14282*, 2023.

[31] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[32] F. Stollenwerk, "Adaptive fine-tuning of transformer-based language models for named entity recognition," *arXiv preprint arXiv:2202.02617*, 2022.

[33] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[34] K. K. Bowden, J. Wu, S. Oraby, A. Misra, and M. Walker, "Slugnerds: A named entity recognition tool for open domain dialogue systems," *arXiv preprint arXiv:1805.03784*, 2018.

[35] M. Ju, M. Miwa, and S. Ananiadou, "A neural layered model for nested named entity recognition," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1446–1459, 2018.

[36] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: the 90% solution," in *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60, 2006.

[37] OpenAI, *Fine-tuning*. OpenAI, 2025. Accessed: 2025-03-26.

[38] OpenAI, "Fine-tuning with direct preference optimization: Step-by-step guide." https://cookbook.openai.com/examples/fine_tuning_direct_preference_optimization_guide, 2024. Accessed 28 Jun 2025.

[39] J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975.

[40] J. Rajamäki, "Ethics of cybersecurity in digital healthcare and well-being of elderly at home," in *Proceeding of the 20th European Conference on Cyber Warfare and Security ECCWS 2021*, Academic Conferences International, 2021.

[41] F. Tronnier, S. Pape, S. Löbner, and K. Rannenberg, "A discussion on ethical cybersecurity issues in digital service chains," in *Cybersecurity of digital service chains: challenges, methodologies, and tools*, pp. 222–256, Springer International Publishing Cham, 2022.

[42] H. Kang, G. Liu, Q. Wang, L. Meng, and J. Liu, "Theory and application of zero trust security: A brief survey," *Entropy*, vol. 25, no. 12, p. 1595, 2023.