Evaluating Zero-Shot and One-Shot Adaptation of Small Language Models in Leader-Follower Interaction

Rafael R. Baptista[1], André de Lima Salgado[2], Ricardo V. Godoy[1], Marcelo Becker[1], Thiago Boaventura[1], and Gustavo J. G. Lahr[3]

***Abstract*— Leader–follower interaction is an important paradigm in human–robot interaction (HRI). Yet, assigning roles in real-time remains challenging for resource-constrained mobile and assistive robots. While large language models (LLMs) have shown promise for natural communication, their size and latency limit on-device deployment. Small language models (SLMs) offer a potential alternative, but their effectiveness for role classification in HRI has not been systematically evaluated. In this paper, we present a benchmark of SLMs for leader–follower communication, introducing a novel dataset derived from a published database and augmented with synthetic samples to capture interaction-specific dynamics. We investigate two adaptation strategies: prompt engineering and fine-tuning, studied under zero-shot and one-shot interaction modes, compared with an untrained baseline. Experiments with Qwen2.5-0.5B reveal that zero-shot fine-tuning achieves robust classification performance (86.66% accuracy) while maintaining low latency (22.2 ms per sample), significantly outperforming baseline and prompt-engineered approaches. However, results also indicate a performance degradation in one-shot modes, where increased context length challenges the model's architectural capacity. These findings demonstrate that fine-tuned SLMs provide an effective solution for direct role assignment, while highlighting critical trade-offs between dialogue complexity and classification reliability on the edge.**

I. INTRODUCTION

As human-robot interaction (HRI) evolves in healthcare and assistive domains, systems are transitioning from executing simple commands to managing complex dyadic coordination [1]–[3]. The primary goal of assistive robotics is to leverage user autonomy. However, effective assistance often requires the robot to intervene or take initiative when the user needs guidance or orientation. In scenarios such as indoor navigation in hospitals or clinics, the robot may need to guide the person to a location or be asked to accompany the person to help her when they reaches the desired location. Additionally, such systems can support remote homecare solutions, particularly in rural areas, for example for elderly coffee farmers in Brazil. In either case, the interaction is effectively modeled through the leader–follower paradigm [4]. In these contexts, role specialization enables the dyad to achieve superior task performance, provided that the robot can accurately determine when to lead, providing direction or proactive support, and when to follow, preserving the user's agency and intended pace [5], [6]. Once classified, this role serves as a high-level trigger for the robot's underlying control architecture, determining whether to execute a proactive guidance policy or a reactive, impedance-based following policy.

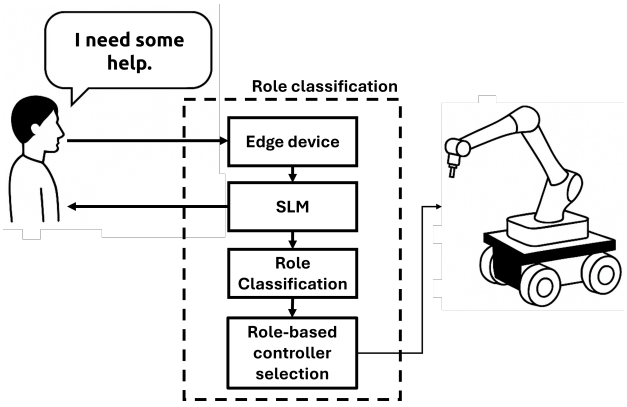


Fig. 1: Edge-deployed leader–follower classification for assistive dyadic interaction. The system architecture selects the most appropriate low-level controller based on the social interation.

Establishing a robust communication channel for this mixed-initiative dialogue is essential to avoid coordination breakdowns and ensure therapeutic or supportive success [7]. Natural language has emerged as a preferred interface for human-robot voice dialogue, as it reduces the unnatural feel of the interaction and increases user acceptance by allowing for fluid, conversational communication [8], [9]. While Large Language Models (LLMs) have shown significant promise in interpreting the linguistic ambiguity and context-dependent nature of such requests [10], [11], their deployment on mobile assistive platforms is frequently hindered by high latency, significant power consumption, and a reliance on stable internet connectivity [12].

These constraints necessitate the adoption of Small Language Models (SLMs) optimized for edge computing. Unlike their larger counterparts, SLMs offer the benefits of being internet-independent and significantly faster, maintaining the real-time responsiveness required for safe and natural HRI without the prohibitive resource demands of massive architectures [13], [14].

Studies have assessed how SLMs perform on resource-constrained edge platforms, highlighting both their feasibility and limitations. Measurement and empirical analyses show that only small models (sub-4B parameters) run reliably

[1]Rafael, Ricardo, Marcelo, and Thiago are with the University of Sao Paulo, Brazil.

[2]André is with Federal University of Lavras, Brazil.

[3]Gustavo is with the Faculdade Israelita de Ensino e Pesquisa Albert Einstein, Hospital Israelita Albert Einstein, Brazil.

on-device, with latency, memory, and thermal constraints severely limiting larger deployments; compression helps but often at the cost of quality [12], [14]. Systems-level research has further explored inference optimizations, such as speculative decoding and sparse mixture-of-experts, to mitigate latency and memory demands on embedded hardware [15], [16]. Application-driven comparisons on NVIDIA Jetson platforms also report that while Qwen2.5-3B is preferred when resources allow [17], Qwen2.5-0.5B remains viable under tighter budgets, showing that carefully selected SLMs can deliver acceptable interactive performance onboard [17], [18]. Collectively, these studies confirm that SLMs can support real-time deployment on mobile robots, but also reinforces the persistent capacity–efficiency tensions of edge inference.

Despite these advances, the literature reveals important gaps in the assignment of leader–follower roles. Most edge-SLM studies emphasize general natural language processing (NLP) or systems metrics (throughput, memory, energy) rather than interaction-specific outcomes such as role inference fidelity, clarification behavior, or task success in closed-loop robot control [12], [14]. Moreover, to the best of our knowledge, there are no publicly available datasets specifically designed for leader–follower communication to support reproducible benchmarking of SLMs on embedded platforms, a gap also emphasized in the systematic review conducted by Corradini et al. [19].

Finally, while zero-shot and one-shot prompting strategies have been extensively studied in NLP [20], [21], their implications for role assignment in leader–follower HRI remain underexplored. Prior work further demonstrates that task-specific prompting can outperform general-purpose approaches [22], [23], highlighting the need to compare prompting strategies with fine-tuning in this domain.

In this work, we address these gaps by conducting a systematic evaluation of an SLM within the leader–follower HRI context, as shown in Fig. 1. Our contributions are twofold: (1) we introduce and make publicly available novel datasets tailored to leader–follower communication in HRI, enabling systematic training and evaluation of models; and (2) we perform a comparative study of prompt engineering and fine-tuning strategies for SLM adaptation on edge devices, explicitly contrasting zero-shot and one-shot prompting with fine-tuned approaches to analyse their respective strengths and limitations.

II. METHODS

This section details the methodological framework adopted in our study. We built our method on processes of synthetic data generation to augment our dataset [24]–[26] and experimental comparisons between prompt engineering and fine-tuning [27]. In the remainder of this section, we first describe the construction of a leader–follower dataset derived from existing dialogue resources and augmented with synthetic samples. Next, we present the rationale for selecting Qwen2.5-0.5B as the base SLM and outline the interaction modes considered (zero-shot and one-shot). We then introduce two adaptation strategies: prompt engineering, where we design zero-shot and one-shot prompt architectures, and fine-tuning, where the model is further specialized for binary role classification.

A. Dataset

As highlighted in the introduction, we conducted extensive searches but found no publicly available datasets specifically tailored to leader–follower communication in HRI[1]. To address this gap, we created our own dataset[2] by leveraging samples from the dialogue and intent classification dataset DailyDialog [28]. We extracted 415 questions that aligned with leader–follower dynamics. The selection was based on identifying questions that required guidance or orientation, and each label was assigned according to the nature of the dialogue in the original context. If one of the participants in the dialogue requested guidance or directions, the label assigned was `LEADER`. Conversely, if a participant initiated a task alone or suggested being accompanied, the label assigned was `FOLLOWER`.

To augment the dataset, we utilized 315 selected questions (148 `FOLLOWER` and 167 `LEADER`), while reserving 100 original human-sourced questions (50 `FOLLOWER` and 50 `LEADER`) for model evaluation. The data augmentation process was conducted using three different LLMs (DeepSeek, Gemini, and GPT-4) following methods established in recent research on synthetic data generation with LLMs [24]–[26], [29]. Each model was instructed to generate approximately six paraphrases per question, preserving the original samples' meaning and structure [30]. This resulted in a total of 5,400 leader-follower-oriented questions, 1,800 per model. This dataset size is consistent with related works that employed a similar number of samples [31], [32].

Two dataset configurations were derived to align with the interaction modes tested in this study. The first corresponds to a *zero-shot* setup and contains leader–follower questions with their associated role labels. The second corresponds to a *one-shot* setup and simulates more complex interactions. It includes the generated questions, an additional clarifying question about the user's intentions, and a response generated by an LLM simulating a human-like but intentionally ambiguous reply. This "scarecrow" validation methodology [33], replaces the need for a human interlocutor during early-stage validation by acting as a temporary black-box LLM component within the system pipeline. This design enables scalable and reproducible experimentation. Moreover, it mitigates ethical risks associated with exposing real participants to potentially misaligned or unvalidated behaviors, functioning as an intermediate validation layer before transitioning to human-subject studies [34].

To evaluate the quality and semantic integrity of the augmented datasets, a semantic fidelity analysis was conducted by comparing the synthetic samples against the original human-annotated data. We employed the Sentence-BERT

[1]Searches in February 2026.

[2]Datasets and prompts are available in our GitHub repository.

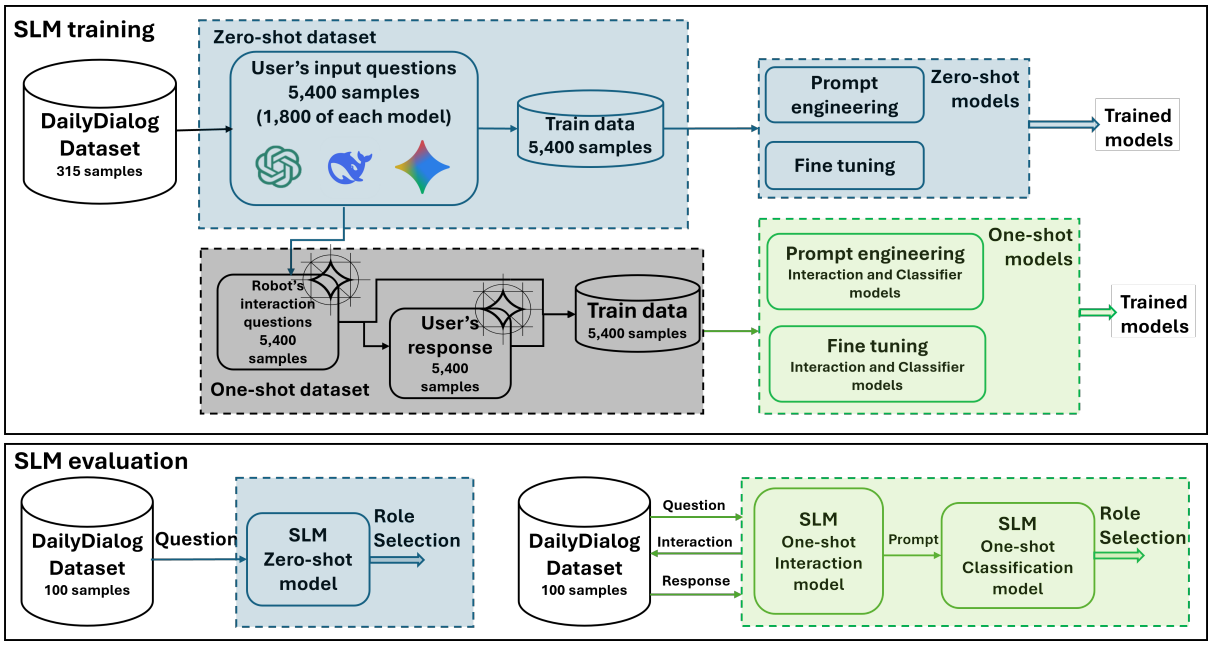


Fig. 2: Overview of the dataset construction and model evaluation pipeline. The top section illustrates the creation of zero-shot and one-shot datasets, generated from DailyDialog samples and synthetic augmentation. The bottom section shows how these datasets were used to train and evaluate models under two adaptation strategies (prompt engineering and fine-tuning) for both zero-shot and one-shot interaction modes.

(SBERT) framework to generate high-dimensional embeddings for both real and synthetic instances [35]. For each synthetic sentence within a specific class, we calculated the maximum cosine similarity relative to all real samples in the same category. This approach ensures a balanced assessment of how well the synthetic generation preserves the core semantic meaning of the original classes, with results interpreted based on the distribution of maximum similarity scores. The mean maximum cosine similarity for the GPT-4 augmented dataset was 0.831 ± 0.086 with respect to the *Follower Rule* and 0.846 ± 0.090 with respect to the *Leader Rule*. For the Gemini-augmented dataset, the corresponding values were 0.786 ± 0.107 and 0.788 ± 0.120, respectively. The DeepSeek-augmented dataset achieved 0.826 ± 0.097 under the Follower Rule and 0.828 ± 0.113 under the Leader Rule.

Finally, each interaction in both datasets was labeled with the correct role assignment, establishing a ground truth for evaluation. The overall process of dataset construction and organization is summarized in Fig. 2. To our knowledge, this represents the first dataset specifically tailored to leader–follower HRI communication, enabling systematic benchmarking of SLM adaptation strategies.

B. Language Model Evaluation

Based on the constraints outlined in the introduction, we required an SLM that balanced computational efficiency with sufficient accuracy for leader–follower role assignment. The literature shows that compact models from the Qwen2.5 family achieve strong parameter efficiency, with the 0.5B variant delivering competitive comprehension and instruction-following performance despite its reduced scale [14]. Benchmark results reported in the Qwen2.5 technical report further demonstrate that this variant outperforms other models of comparable size across standard evaluations while maintaining low latency, reinforcing its suitability for deployment on resource-constrained robotic platforms [17]. These findings support our choice of Qwen2.5-0.5B as the base model. Its balance of computational efficiency and task performance makes it well-suited for embedded assistive robotics, where real-time responsiveness and resource efficiency are critical.

We evaluated this model under two *interaction modes*: zero-shot and one-shot. In zero-shot mode, the SLM must classify the leader–follower role using only the user's initial input, with no opportunity for clarification. In the one-shot mode, the SLM is permitted to ask a single clarifying question, after which it combines the original input and the user's response to make its prediction. The bottom half of Fig. 2 illustrates how these datasets are integrated into the evaluation pipeline, where zero-shot and one-shot models are trained and assessed using prompt engineering and fine-tuning strategies. Together, these modes provide a framework for systematically evaluating the impact of clarification on role assignment. The performance of all proposed strategies was evaluated using Monte Carlo Cross-Validation (MCCV), with 30 independent iterations to ensure statistical de-biasing. We next describe the baseline and the two adaptation strategies applied to this model: prompt engineering and fine-tuning. As a reference condition, the Qwen2.5-0.5B model, without any task-specific adaptation, was used as baseline.

TABLE I: Performance comparison of evaluated models across zero-shot and one-shot interaction modes. Metrics include classification accuracy, precision, recall, and F1-score, alongside efficiency measures (tokens per second and latency).

Interaction	Method	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	Tokens/s	Latency [ms]
Zero-shot	**Baseline**	55.00±10.98	75.17±21.61	16.39±6.53	25.99±8.23	281.9±105.8	33.8±3.7
	Prompt Eng.	53.87±10.38	56.80±19.20	31.03±12.58	38.80±12.66	85.9±27.0	111.9±40.0
	Fine-tuning	86.66±6.77	84.81±10.07	90.05±7.98	86.72±6.53	432.1±148.2	22.2±4.2
One-shot	**Baseline**	48.00±9.09	40.83±32.16	9.27±7.99	14.18±11.39	988.63±116.7	41.0±2.6
	Prompt Eng.	45.07±8.17	42.48±26.49	17.41±10.91	23.12±13.42	305.6±90.7	213.8±309.1
	Fine-tuning	51.65±13.40	43.69±45.02	8.90±13.29	13.07±16.55	1851.69±380.09	22.2±3.5

1) Baseline Model: We employed Qwen2.5-0.5B model with its original weights as the baseline, without applying few-shot examples or any task-specific adaptation. The model received only the task instruction and generated responses based on its pre-trained parameters.

2) Prompt Engineering: We designed prompt architectures for both the zero-shot and one-shot setups. The zero-shot architecture employs a single system prompt that encapsulates the agent's personality, behavior, and task-specific instructions, drawing on prompt engineering methods such as few-shot prompting and meta-prompting [21], [35], [36].

In contrast, the one-shot architecture uses two system prompts, each tailored to a specific functional component of the agent. The first prompt is designed to elicit a clarifying question from the model. In contrast, the second integrates both the user's initial input and their clarification response to determine the final role assignment. This decomposition was motivated by the limitations of small-scale models, where breaking instructions into smaller, focused prompts can improve consistency and overall performance [21], [36].

3) Fine-tuning: While prompt engineering can adapt an SLM to specific tasks, its effectiveness is often constrained by the model's pretraining distribution, especially in small-scale models [37], [38]. To address this limitation, we fine-tuned the Qwen2.5-0.5B model on our generated dataset to specialize it for binary leader–follower role classification. Fine-tuning has been widely adopted to improve task adherence when zero-shot or few-shot prompting alone proves insufficient. The fine-tuning process was conducted on an NVIDIA server equipped with L40S 48GB GPUs.

For this step, we used the Autotrain framework for text classification [39], which we chose for its practicality and efficiency in hyperparameter exploration. In this framework, the base model is adapted for the classification task by appending a lightweight linear head to the pretrained architecture. This head learns to map the model's latent representations into discrete labels, bridging the gap between general-purpose language understanding and the specific requirements of leader–follower classification. The parameters used on the framework are: 10 training epochs, a learning rate warmup of 50 steps, weight decay of 0.01, and a batch size of 16 for both training and evaluation. The model was optimized using mixed-precision (FP16) training, with the best model selected based on accuracy and a stratified 20% validation split.

In the zero-shot condition, a single fine-tuned model is sufficient to directly classify the role based on the user's initial request. In the one-shot condition, however, we trained two complementary models: one fine-tuned to generate a clarifying question when the initial input is ambiguous, and a second fine-tuned to perform the final leader–follower classification using both the original input and the user's follow-up answer. This separation ensures that the clarification stage is optimized for interaction quality, while the classification stage remains optimized for decision accuracy.

III. EXPERIMENTAL EVALUATION

This section presents the evaluation outcomes of the proposed leader–follower classification framework. We first report the overall performance of different adaptation strategies (baseline, prompt engineering, and fine-tuning) across zero-shot and one-shot interaction modes. Each proposed SLM model was evaluated on the test set of 100 phrases, reporting key performance metrics such as overall accuracy, per-class precision, recall, and F1-score, together with system efficiency in terms of throughput and latency. To ensure statistical robustness, this evaluation was repeated thirty times, allowing us to run statistical analysis on the results. Experiments were conducted under both zero-shot and one-shot conditions. Based on these results, we identify the best-performing model and, in a subsequent subsection, perform a focused analysis on the impact of sentence length on classification performance.

A. Classification Metrics

Table I summarizes the results across methods and interaction modes. For the zero-shot interaction, fine-tuning significantly outperformed both the baseline and prompt engineering strategies. In this setup, fine-tuning achieved an accuracy of 86.66%, compared to 53.87% for prompt engineering and 55.00% for the baseline (z-test, $p < 0.001$). A different pattern was observed in the one-shot setup, where performance across all methods converged near chance level; fine-tuning reached only 51.65% accuracy, statistically comparable to the baseline (48.00%) and prompt engineering (45.07%) (z-test, $p > 0.001$). Precision and recall metrics reveal that while zero-shot fine-tuning achieves a robust balance, the zero-shot baseline and prompt-engineering methods are significantly

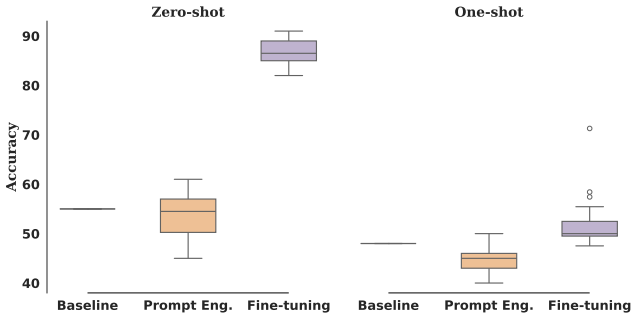


Fig. 3: Accuracy dispersion across 30 runs for each interaction mode and adaptation strategy. Fine-tuned models consistently achieve higher median accuracy than baseline and prompt-engineered approaches, though the one-shot fine-tuned condition shows slightly greater variability.

skewed toward higher precision at the expense of extremely low recall (e.g., 16.39% for the baseline). This suggests that without specialization, the models are overly conservative, failing to detect the majority of leader–follower transitions. In the one-shot configuration, however, fine-tuning failed to maintain this performance, with recall dropping to a marginal 8.90%. This shift indicates that, while fine-tuned SLMs are highly effective for direct classification, the increased semantic complexity and context length of one-shot interactions—incorporating both the clarifying question and the simulated response—may exceed the 0.5B model's limited parameter capacity, leading to a breakdown in classification.

When comparing zero-shot and one-shot modes, the fine-tuned model's performance degraded significantly, with accuracy dropping from 86.66% to 51.65% ($p < 0.001$). Similarly, the baseline accuracy declined from 55.00% to 48.00%, while prompt engineering results decreased from 53.87% to 45.07%. The dispersion observed in Fig. 3 suggests that the one-shot interaction mode introduces a level of linguistic complexity that the 0.5B parameter model struggles to process reliably. This is most evident in the fine-tuned one-shot metrics, where precision fell to 43.69% and recall collapsed to a marginal 8.90%. These results indicate that while fine-tuned SLMs are highly effective for direct, single-turn role classification, the increased context length—incorporating both a clarifying question and a synthetic response—likely exceeds the model's capacity to maintain semantic fidelity. In leader–follower HRI, this is an important finding: it suggests that for small models on the edge, a more complex, multi-turn dialogue may actually increase the risk of coordination breakdowns by introducing noise that the model cannot successfully filter to determine initiative. Further research into specialized context-pruning pipelines is necessary to enable the linguistic flexibility of multi-turn dialogues without compromising the system's initiative-detection reliability.

Finally, efficiency analysis revealed clear trade-offs between adaptation strategies. Fine-tuning consistently achieved the lowest latency (around 22 ms per sample) and the highest throughput (432–1851 tokens/s), significantly outperforming both baseline and prompt engineering. Prompt engineering incurred the highest latency, particularly in the one-shot mode (213.8 ms), due to the added prompt complexity and the need to maintain longer context windows within the 0.5B model's inference cycle. Interestingly, while the one-shot configuration produced more tokens on average—due to the interaction including both a clarifying question and a synthetic response—the measured latency for the fine-tuned model remained identical to the zero-shot case (22.2 ms). Due to high throughput, fine-tuned one-shot models incur negligible latency overhead compared to zero-shot modes. However, the trade-off is not negligible for prompt-engineered approaches, where latency increased nearly twofold in one-shot mode.

B. Sentence Length Analysis

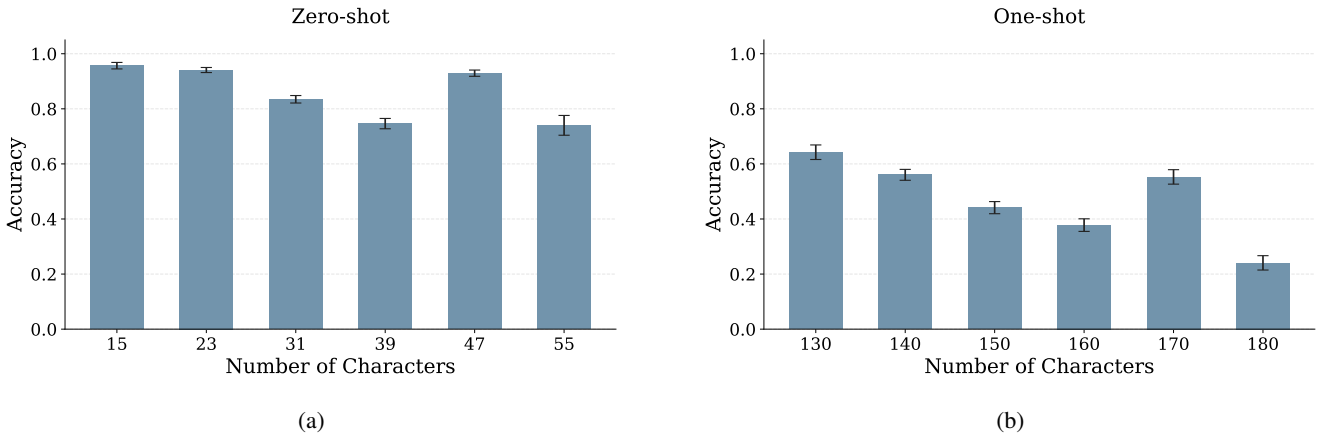
Given recent reports that longer inputs can exacerbate language model hallucinations [40], we evaluated whether sentence length affects the reliability of leader–follower classification. Specifically, we tested whether increased context in longer inputs improved decision quality or increased the model's susceptibility to misclassification. To analyze this, we performed data binning on the test set with a step size of 10 characters, calculating the mean accuracy and standard deviation across thirty independent runs for each bin. The x-axis of Fig. 4 indicates the character length bins, while the y-axis represents the classification accuracy.

The results reveal distinct patterns for the two interaction modes. In the zero-shot configuration (Fig. 4a), accuracy remained robust across most bins, maintaining levels above 80% for shorter sentences and showing relative stability even as length increased, despite a localized dip in the 39-character bin. This suggests that the 0.5B model effectively captures role intent in single-turn interactions regardless of moderate phrasing complexity. By contrast, the one-shot configuration (Fig. 4b) exhibited a significant and progressive decline in performance. Accuracy dropped from approximately 65% in the shortest bin to nearly 25% for the longest sentences.

Taken together, these findings indicate that while zero-shot classification is relatively resilient to sentence length, one-shot interactions struggle with the combination of longer inputs and multi-turn context leads to a breakdown in semantic fidelity. For extremely small models on the edge, the increased noise of a verbose, multi-turn dialogue appears to overwhelm the model's parameters, underscoring the importance of concise interaction design when using sub-1B parameter models for initiative-switching.

IV. CONCLUSION

This work presented an evaluation of SLMs for leader–follower role classification in HRI. We constructed a novel dataset tailored to this task, introduced two interaction modes (zero-shot and one-shot), and compared prompt engineering and fine-tuning strategies on the Qwen2.5-0.5B model. Across thirty independent trials, zero-shot fine-tuning



(a) (b)

Fig. 4: Accuracy as a function of sentence length for the fine-tuned models under (a) zero-shot and (b) one-shot settings. Error bars indicate one standard deviation. For the zero-shot setting, sentence length is measured using only the user's question. In the one-shot setting, sentence length is computed from the concatenation of the user's question, the provided example answer, and the final response, resulting in a larger number of characters overall.

consistently outperformed prompt engineering and baseline methods in both accuracy and efficiency. However, one-shot interactions exhibited a significant performance degradation, highlighting critical architectural limits when processing multi-turn context on extremely small models. Together, these results reveal trade-offs among efficiency, precision, and dialogue complexity when deploying SLMs on edge-constrained robotic platforms.

Beyond empirical performance, our study makes two main contributions to the field. First, we provide the first publicly available dataset designed specifically for leader–follower HRI communication, enabling reproducible benchmarking of adaptation strategies. Second, we demonstrate that fine-tuning is not only superior in accuracy but also more efficient in terms of latency and throughput, directly addressing the practical constraints of embedded robotic platforms. In this context, our sentence length analysis revealed that while one-shot configurations are less sensitive to input complexity and length than zero-shot models, they require greater parameter capacity or specialized context management to maintain classification fidelity.

Despite the observed poor performance in 0.5B models, one-shot interactions retain significant potential as they closely mimic the inherent complexity of real-life dyadic coordination. To transition from direct role assignment to these more natural dialogue paradigms, future research must develop specialized evaluation frameworks for the edge. Such frameworks are needed to assess the semantic quality of the SLM's clarifying questions and the contextual veracity of the scarecrow simulation responses. This necessitates a shift toward either rigorous mathematical evaluations of semantic fidelity or the inclusion of subjective user judgment through human-in-the-loop studies.

There remain, however, important directions for future work. Our dataset, while novel, is synthetic and limited in scale; extending it with larger, multimodal, and human-annotated samples will strengthen the ecological validity of the findings. Similarly, exploring alternative SLM architectures, compression strategies, and multilingual settings may further improve efficiency and inclusion. Finally, evaluating these models in real-world robotic experiments, where communication is embedded in dynamic tasks and environments, will be critical to validate their robustness and user acceptance. Future research will also address the underexplored potential of one-shot interactions by conducting deeper evaluations of the SLM-generated answers and the simulation (scarecrow) module's performance, using both automated metrics and human-in-the-loop studies to refine multi-turn dialogue reliability.

ACKNOWLEDGMENT

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil, process Number 2025/09390-3, and by the Hospital Israelita Albert Einstein's Seed Money 2024 call. We also acknowledge Minas Gerais State Research Support Foundation (FAPEMIG) - APQ-00996-24; *CAPES*: Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES), via institutional funding (Code 001); UFLA – Universidade Federal de Lavras, funding N 014/2026 – INOVAÇÃO process Number 23090.000597/2026-73; 2024 Cultural Technology Research and Development Project by the Ministry of Culture, Sports and Tourism and the Korea Creative Content Agency(Project Name: Development of AX Authoring Technology and Immersive Content Based on MR Spatial Computing, and Global Specialist Training, Project Number: RS-2024-00399186, Contribution Rate: 100%); and *CNPq*: National Council for Scientific and Technological Development (CNPq), for partial financial support. Finally, this work was also supported by the Petróleo Brasileiro S/A - Petrobras, using resources from the R&D clause of the ANP, in partnership with the Universidade de São Paulo (USP) and the Fundação de Apoio à Física e à Química (FAFQ), under Cooperation Agreement No. 2023/00016-6 and 2023/00013-7.

REFERENCES

[1] A. Asadi, D. Herath, G. Shaw, G. Caldwell, and E. Williams, "Redrawing boundaries: Systemic impacts of rehabilitation robots in clinical care settings," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 1216–1220.

[2] S. Adebayo, J. C. Dessing, and S. McLoone, *Dyadic Human-Robot Interaction: Emerging Technologies, Challenges, and Opportunities*. Singapore: Springer Nature Singapore, 2026, pp. 1–51.

[3] J. Tschida, K. Michael, and T. McDaniel, "Exploring social robots for healthy older adults: Aging with companionship," *IEEE Transactions on Technology and Society*, 2025.

[4] A. Takai, Q. Fu, Y. Doibata, G. Lisi, T. Tsuchiya, K. Mojtahedi, T. Yoshioka, M. Kawato, J. Morimoto, and M. Santello, "Role specialization enables superior task performance by human dyads than individuals," *The International Journal of Robotics Research*, 8 2025.

[5] A. Noormohammadi-Asl, K. Fan, S. L. Smith, and K. Dautenhahn, "Human leading or following preferences: Effects on human perception of the robot and the human–robot collaboration," *Robotics and Autonomous Systems*, vol. 183, p. 104821, 2025.

[6] M. Kwon, M. Li, A. Bucquet, and D. Sadigh, "Influencing leading and following in human-robot teams." in *Robotics: Science and Systems*, 2019.

[7] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, vol. 1, no. 4, pp. 279–285, 2020.

[8] A. Zubiaga, "Natural language processing in the era of large language models," *Frontiers in Artificial Intelligence*, vol. 6, 1 2024.

[9] N. Cherakara, F. Varghese, S. Shabana, N. Nelson, A. Karukayil, R. Kulothungan, M. Afil Farhan, B. Nesset, M. Moujahid, T. Dinkar, V. Rieser, and O. Lemon, "FurChat: An embodied conversational agent using LLMs, combining open and closed-domain dialogue with facial expressions," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, Eds. Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 588–592.

[10] C. Wang, S. Hasler, D. Tanneberg, F. Ocker, F. Joublin, A. Ceravola, J. Deigmoeller, and M. Gienger, "Lami: Large language models for multi-modal human-robot interaction," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–10.

[11] J. Lim, S. Patel, A. Evans, J. Pimley, Y. Li, and I. Kovalenko, "Enhancing human-robot collaborative assembly in manufacturing systems using large language models," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2024, pp. 2581–2587.

[12] N. Dhar, B. Deng, D. Lo, X. Wu, L. Zhao, and K. Suo, "An empirical analysis and resource footprint study of deploying large language models on edge devices," in *Proceedings of the 2024 ACM southeast conference*, 2024, pp. 69–76.

[13] H. Kim, H. Hwang, J. Lee, S. Park, D. Kim, T. Lee, C. Yoon, J. Sohn, J. Park, O. Reykhart, T. Fetherston, D. Choi, S. H. Kwak, Q. Chen, and J. Kang, "Small language models learn enhanced reasoning skills from medical textbooks," *npj Digital Medicine*, vol. 8, no. 1, p. 240, May 2025.

[14] X. Yan and Y. Ding, "Are we there yet? a measurement study of efficiency for llm applications on mobile devices," in *Proceedings of the 2nd International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things*, 2025, pp. 19–24.

[15] D. Xu, W. Yin, H. Zhang, X. Jin, Y. Zhang, S. Wei, M. Xu, and X. Liu, "Edgellm: Fast on-device LLM inference with speculative decoding," *IEEE Transactions on Mobile Computing*, vol. 24, no. 4, pp. 3256–3273, 2025.

[16] R. Yi, L. Guo, S. Wei, A. Zhou, S. Wang, and M. Xu, "Edgemoe: Empowering sparse large language models on mobile devices," *IEEE Transactions on Mobile Computing*, vol. 24, no. 8, pp. 7059–7073, 2025.

[17] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[18] M. A. J. Gómez, R. Jiménez-Moreno, and A. A. Espitia-Cubillos, "Tiny large language models in embedded NVIDIA portable hardware: Comparative analysis," *Journal of Hunan University (Natural Sciences)*, vol. 52, no. 4, pp. 12–24, 2025.

[19] F. Corradini, M. Leonesi, and M. Piangerelli, "State of the art and future directions of small language models: A systematic review," *Big Data and Cognitive Computing*, vol. 9, no. 7, p. 189, 2025.

[20] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, "An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing," *JMIR Medical Informatics*, vol. 12, no. 1, 2024.

[21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[22] M. G. Ozsoy, "Multilingual prompts in llm-based recommenders: Performance across languages," *arXiv preprint arXiv:2409.07604*, 2024.

[23] J. Mendonça, P. Pereira, H. Moniz, J. P. Carvalho, A. Lavie, and I. Trancoso, "Simple llm prompting is state-of-the-art for robust and multilingual dialogue evaluation," *arXiv preprint arXiv:2308.16797*, 2023.

[24] M. Nadas, L. Diosan, and A. Tomescu, "Synthetic Data Generation Using Large Language Models: Advances in Text and Code," *IEEE Access*, vol. 13, pp. 134 615–134 633, 2025.

[25] J. Ren, Z. Du, Z. Wen, Q. Jia, S. Dai, C. Wu, and Z. Dong, "Few-shot LLM Synthetic Data with Distribution Matching," Feb. 2025.

[26] M. Josifoski, M. Sakota, M. Peyrard, and R. West, "Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction," Oct. 2023.

[27] X. Zhang, N. Talukdar, S. Vemulapalli, S. Ahn, J. Wang, H. Meng, S. M. B. Murtaza, D. Leshchiner, A. A. Dave, D. F. Joseph, M. Witteveen-Lane, D. Chesla, J. Zhou, and B. Chen, "Comparison of Prompt Engineering and Fine-Tuning Strategies in Large Language Models in the Classification of Clinical Notes," *medRxiv*, p. 2024.02.07.24302444, Feb. 2024.

[28] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[29] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang, "On llms-driven synthetic data generation, curation, and evaluation: A survey," *arXiv preprint arXiv:2406.15126*, 2024.

[30] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, L. A. Tuan, and S. Joty, "Data augmentation using llms: Data perspectives, learning paradigms and challenges," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 1679–1705.

[31] D. Grießhaber, J. Maucher, and N. T. Vu, "Fine-tuning bert for low-resource natural language understanding via active learning," *arXiv preprint arXiv:2012.02462*, 2020.

[32] N. Alex, E. Lifland, L. Tunstall, A. Thakur, P. Maham, C. J. Riedel, E. Hine, C. Ashurst, P. Sedille, A. Carlier *et al.*, "Raft: A real-world few-shot text classification benchmark," *arXiv preprint arXiv:2109.14076*, 2021.

[33] T. Williams, C. Matuszek, R. Mead, and N. Depalma, "Scarecrows in oz: the use of large language models in hri," pp. 1–11, 2024.

[34] R. Wen, F. Ferraro, and C. Matuszek, "Gpt-4 as a moral reasoner for robot command rejection," in *Proceedings of the 12th International Conference on Human-Agent Interaction*, 2024, pp. 54–63.

[35] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[36] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 3816–3830.

[37] M. Gao, X. Hu, X. Yin, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *Computational Linguistics*, pp. 1–27, 2025.

[38] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[39] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," *arXiv preprint arXiv:2408.13296*, 2024.

[40] N. Chakraborty, M. Ornik, and K. Driggs-Campbell, "Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–35, 2025.