An Empirical Study on Improving an LLMs based Interviewer against Human Interviewees

Rikuto Tsuchida, Ryo Hasegawa, Tetsuya Sonoda, Takehito Utsuro

Graduate School of Science and Technology, University of Tsukuba

{s2520796,s2420791,s2630267}_@u.tsukuba.ac.jp,
utsuro_@iit.tsukuba.ac.jp

Abstract

This paper presents a semi-structured interview system composed of multiple LLM components, in which an LLM-based interviewer conducts dialogue and another LLM estimates interviewee persona attributes from the dialogue history and dynamically updated slot information. The system prioritizes questions related to important yet unconfirmed attributes to support efficient interviews. We evaluate persona attribute estimation performance when interviewing both human participants and LLM-based pseudo interviewees under zero-shot and few-shot prompting conditions. Experiments across three domains (IT engineer, store staff, and school teacher) with a total of 39 interviews show that persona attribute estimation is more challenging for human interviewees than for the LLM pseudo interviewees under zero-shot prompting. However, introducing few-shot prompting consistently improves estimation performance for human interviewees, primarily by increasing recall. On the other hand, its effect on the LLM pseudo interviewees is mixed across domains. A case study confirms that persona attribute estimation improves when illustrative examples are provided.

1 Introduction

With the advancement of Large Language Models (LLMs) (Brown et al., 2020), dialogue systems capable of natural conversation with humans are being increasingly utilized in scenarios such as interviews and customer service.

This paper builds on prior work (Hasegawa et al., 2025) that conducted semi-structured interviews (Fielding, 2003a,b; Wengraf, 2001) with an LLM-based user simulator as a pseudo interviewee. We explain semi-structured interviews in Section 3. As discussed in Section 2, that study showed that the proposed system can effectively carry out such interviews with the simulator.

However, the simulator answers from predefined persona settings and replies "I don't know" to questions outside that scope, whereas human interviewees often use vague expressions, include multiple pieces of information in a single utterance, deviate from the intended topic, and rephrase or shift perspective rather than frankly admitting a lack of knowledge (Section 5). These more diverse dialogues are not identical to unconstrained real-world interviews, but they are closer to them than simulator dialogues and therefore provide a more informative test of interview systems. The objective of this paper is to improve the ability to accurately estimate a subject's persona attributes based on the conversation history when conducting interviews with human participants.

In this paper, we use the semi-structured interview approach shown in Figure 1 and conduct interviews with both human interviewees and LLM-based pseudo interviewees.

Section 3 presents the proposed interview system, and Section 4 describes the systems used in this paper and how the LLMs estimate persona attributes and values. In the proposed system, an LLM responsible for estimating the interviewee's persona attributes infers the interviewee's attributes at each dialogue turn based on the dialogue history and information recorded in slots. This functionality enables the system to identify important information among the information obtained through the interview, allowing prioritized presentation of questions related to important persona attributes, thereby achieving more efficient interviews.

In Section 7, to verify the effectiveness of this system for human interviewees, we evaluate the performance of the LLM responsible for estimating persona attributes. Specifically, we evaluate the performance in interviews with both human interviewees and LLM-based pseudo interviewees, and assess the persona attribute estimation performance achieved by introducing few-shot prompting

with examples (Figure 2, Figure 3, Figure 5 in Section A of Appendix). Furthermore, we conduct case analysis on instances where results improved with the introduction of few-shot prompting. The contributions of this paper are as follows:

1. Based on the semi-structured interview framework, we repeated the experiments previously conducted with LLM-based pseudo interviewees, this time involving both human interviewees and LLM-based pseudo interviewees, and compared the system's persona attribute estimation performance in both types of interviews.

2. We introduced zero-shot and few-shot prompts to the LLM performing persona attribute estimation and demonstrated that few-shot prompting improves persona attribute estimation performance in dialogues with human interviewees.

3. We performed case analysis on instances where persona attribute estimation output improved with the introduction of few-shot prompting, and demonstrated cases where few-shot prompting functions effectively.

2 Related Work

In the field of dialogue systems, extensive research has been conducted on interview-oriented dialogue systems. Interview dialogue encompasses diverse elements such as question generation, dialogue management, and dialogue strategies, and has been the subject of wide-ranging research (Zeng et al., 2023; DeVault et al., 2014; Johnston et al., 2013; Kobori et al., 2016; B et al., 2020; Nagasawa et al., 2024; Ge et al., 2023; Inoue et al., 2020). Among these studies, research on dialogue systems for job interviews is notable (Su et al., 2019, 2018). Inoue et al. (2020) developed a system in which an android robot serves as the interviewer. This system analyzes keywords and satisfaction levels contained in user utterances and generates follow-up questions based on four predefined topics. Schatzmann and Young (2009) proposed a statistical modeling approach that simulates user responses for the purpose of optimizing dialogue strategies.

In recent years, rapid advances in LLMs have led to improvements in dialogue system performance (Geiecke and Jaravel, 2024). In particular, combining LLMs with slot-filling architectures has been shown to achieve both advanced language understanding capabilities and flexible dialogue control (Hudeček and Dusek, 2023; Jacqmin et al., 2022; Siddique et al., 2021; Coope et al., 2020). Furthermore, by leveraging the high contextual understanding capabilities of LLMs such as GPT, question generation based on dialogue flow and previous utterance content, as well as flexible interview dialogue according to the situation, have become possible (Sun et al., 2024; Feng et al., 2023; Heck et al., 2023). Additionally, Wagner and Ultes (2024) demonstrated that combining LLMs with rule-based or slot-based dialogue management enables finer control of dialogue flow. Research on dynamic slot generation has also been conducted in fields other than interview systems. For example, Komada et al. (2024) proposed a method for maintaining narrative consistency by automatically generating slots from dialogue history in tabletop role-playing games (TRPGs). In the medical field, Hashimoto et al. (2025) developed a dialogue system to support career interviews by nursing managers and reported improvements in both interview efficiency and quality. Zenimoto et al. (2025) proposed a dialogue system that collects health questionnaire information from elderly people through casual conversation, implementing topic-guided dialogue that acquires target information by guiding LLMs using Chain-of-Thought prompting.

Additionally, Hasegawa et al. (2025) proposed a dialogue control framework using LLMs that can efficiently ask questions by accumulating information obtained during dialogue as slots in semi-structured interviews with multiple interviewees. Based on these prior studies, this research focuses on semi-structured interviews that dynamically generate and update slot sets based on interviewee utterance content. While research on semi-structured interviews has been conducted previously (Parfenova, 2024; Hu et al., 2024), most rely on static or manually designed control strategies, and the flexible generation and control of questions according to subject utterances has not been sufficiently examined. Furthermore, the utilization of information obtained from past interviews in new interviews has not been adequately studied. This study evaluates the proposed framework not only through simulations with LLM-based pseudo interviewees but also through semi-structured interviews with human participants. It further investigates the impact of introducing few-shot prompting on persona attribute estimation based on dialogue history and

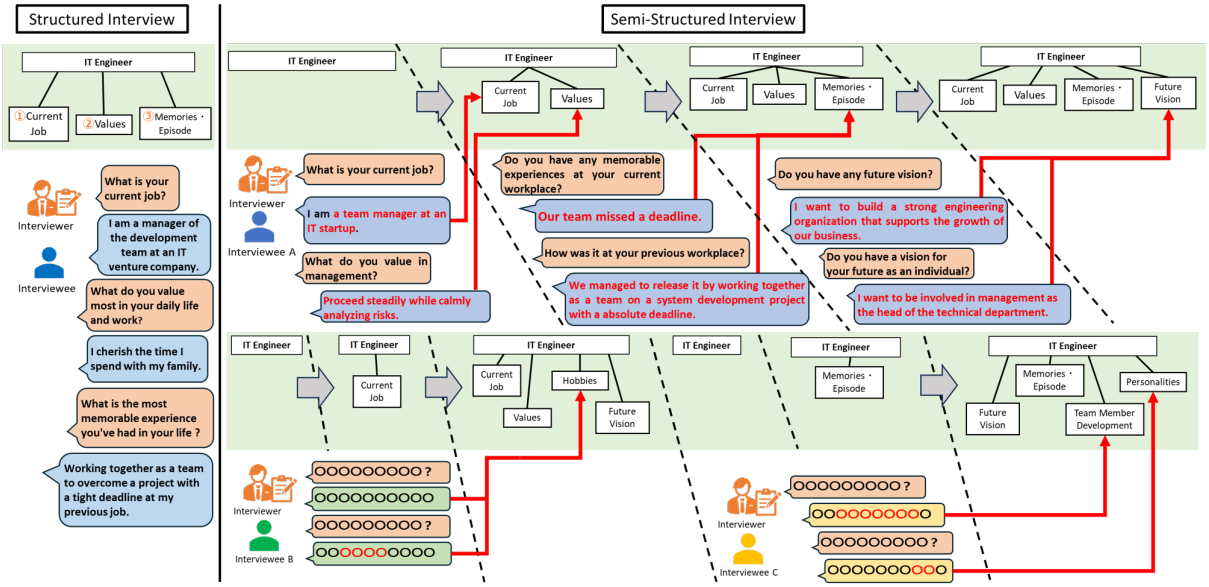

Figure 1: Diagram of Structured and Semi-Structured Interviews

slot information.

3 Semi-Structured Interviewer by LLMs

In this study, we implement a semi-structured interview system using LLMs based on the dialogue control framework (Hasegawa et al., 2025). Semi-structured interviews are an interview format in which the interviewer can flexibly ask additional questions according to the interviewee's response content. Since questions can be adapted to the interviewee and situation, this approach has the advantage of eliciting more diverse responses compared to structured interviews based solely on predetermined questions. The upper half of Figure 3 shows the overall processing flow of the interview system used in this study. The system consists of multiple LLMs that generate questions, create new slots, and record and update slot values from the interviewee's utterances.

Figure 3 illustrates the processing flow of the proposed semi-structured interview system, including LLM modules for question generation and slot creation and update, and shows where each step takes place within the interactive system. In step 1, the system generates a new slot, "Current jobs", and a new question is generated (step 2). Next, the interviewee answers the question (step 3), and the system fills the newly created slot with the answer (step 4). Finally, the system determines whether the saved response qualifies as a persona attribute. If it does, the response is saved with the corresponding slot values filled in (step 5). In other words, step

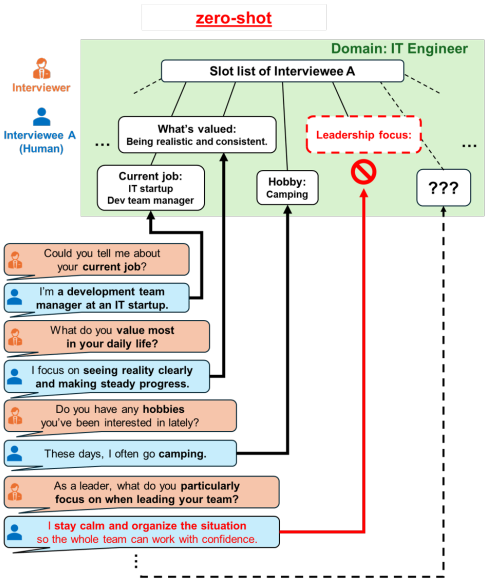

Figure 2: Diagram of Semi-Structured Interviews with Zero-Shot Prompts

3 is the interviewee's answering of the question, whereas step 5 is the system's inference, based on that answer, of whether the interviewee possesses the corresponding persona attributes. The system then determines whether to proceed to the next question; if it determines that the interview has been sufficiently conducted, the interview ends (step 6).

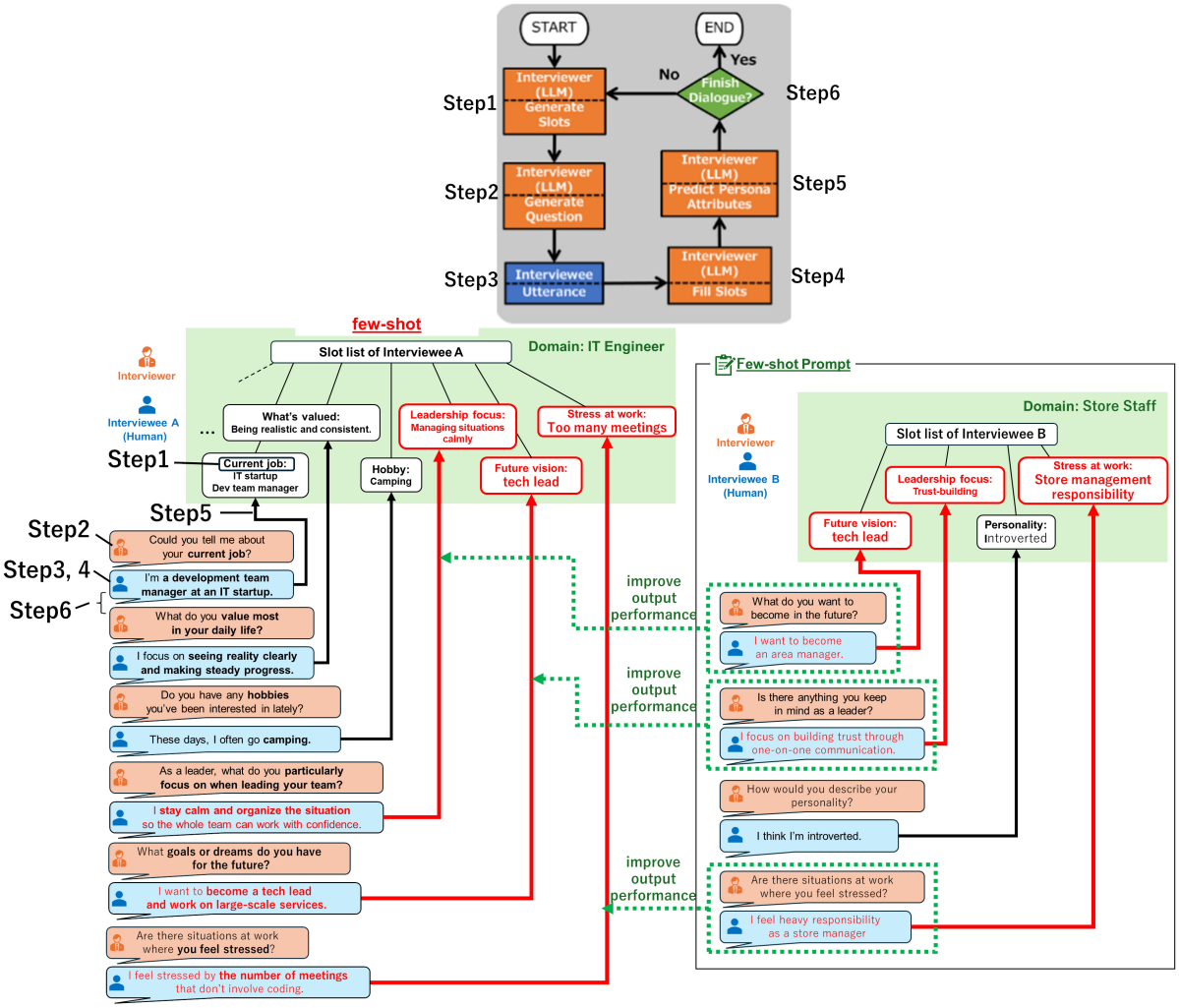

Figure 3: Diagram of Semi-Structured Interviews with Few-Shot Prompts

4 Persona Attributes and Values Estimation through Semi-Structured Interviews

A persona represents the character or role of an interviewee, referring to personal information and background information. Information collected through interviews is maintained as slots within the system. Each slot consists of a slot name and value pair. The slot name represents an abstract topic or attribute category (e.g., current career, hobbies, etc.), and the value stores the interviewee's utterance content. At each turn of the interview, the system generates new slots and updates the values of existing slots based on the dialogue history.

New slot generation is performed through two types of methods. One is a method where the LLM generates slots for asking in-depth questions to the interviewee. The other is a method for generating new slots from a diverse candidate set of slot names. As one component of this system, we introduce an LLM that estimates the interviewee's persona attributes based on dialogue history and slots. This LLM estimates the interviewee's persona attributes at each turn from the dialogue history and currently held slots. The estimation result, the estimated persona attributes, is labeled with whether the question regarding that attribute has not yet been asked or has already been asked. This enables prioritized generation of questions related to unasked estimated persona attributes, avoiding duplicate questions while improving information collection efficiency.

5 Human versus LLMs as Interviewees

There are differences in interview dialogues depending on whether the interviewee is human or an LLM (Figure 5 in Section A of Appendix).

Human interviewees typically provide diverse responses: vague or ambiguous expressions, an-

swers that do not fully align with the question, multiple pieces of information in a single utterance, topic deviation, or a shift in perspective rather than answering "I don't know." In contrast, the LLM-based pseudo interviewee used in this study, as well as that of Hasegawa et al. (2025), generates responses based on predefined persona settings, or replies "I don't know" when the question is outside that scope. We give human participants the same instruction as the LLM interviewee: when a question is not covered by the assigned persona, they are told not to invent an answer. However, unlike the LLM, humans do not restrict themselves to the fixed response "I don't know," and this remaining diversity is a challenge addressed in this paper. As shown in Table 2 in Section A of Appendix, when the interviewee is an LLM-based pseudo interviewee, a specific response other than "I don't know" about "leadership style" led the system to output that attribute. When the interviewee is human, the system did not treat the corresponding answer as important, and the attribute was not output.

6 Few-Shot Prompt for LLM-based Interviewer

In the proposed interview system, we apply few-shot prompting to the LLM responsible for estimating the interviewee's persona attributes (Figure 3). Under the zero-shot condition, the persona attribute estimation LLM's prompt is given no examples at all, and persona attributes are estimated based solely on the interviewee's utterance content. In this case, since the LLM is not provided with explicit decision criteria through examples, it relies primarily on knowledge from its own pretraining and the interviewee's response content to estimate persona attributes. As a result, estimation may become unstable, particularly for diverse and ambiguous utterances from human subjects.

In contrast, under the few-shot condition, a small number of examples for persona attribute estimation are provided as prompts to the persona attribute estimation LLM. This gives the estimation LLM explicit decision criteria for inferring persona attributes. As a result, the system can more appropriately estimate persona attributes from the subject's utterance content, and improvements in estimation accuracy are expected. This comparison is important for clarifying the impact of few-shot prompting on persona attribute estimation performance for both LLM-based pseudo interviewees and human interviewees. Specifically, we examine how few-shot prompting affects estimation for LLM-based pseudo interviewees, and how effectively it handles the diverse utterances of human interviewees.

7 Evaluation

7.1 The Procedure

In this study, we used the proposed interview system as the interviewer and set either an LLM-based pseudo interviewee or a human subject as the interviewee. For all LLMs in the system, we used GPT-4o[1] (gpt-4o-2024-11-20). Both LLM and human interviewees were provided with pre-created persona settings in advance to ensure reproducibility. The persona settings assigned individual values to each of 10 common persona attributes (Table 3 in Section A of Appendix). Personas were classified into three domains: "IT engineer," "store staff," and "school teacher." We prepared 10 types of personas for each domain, and the LLM-based pseudo interviewee was designed to respond to questions related to content described in the persona settings based on the described content, while responding with "I don't know" when asked about topics not described in the persona settings.

For dialogue control implementation, we used LangGraph[2].

First, using zero-shot prompting for the prompt of the LLM responsible for estimating persona attributes, we conducted interviews sequentially with both human and LLM interviewees. We conducted interviews with three LLM-based pseudo interviewees for comparison purposes and 13 human interviewees in each of the three domains. These 39 human participants were all different individuals; each was assigned a distinct persona profile and took part in only one domain, for a total of 9 LLM interviews and 39 human interviews.

Next, using few-shot prompting for the prompt of the LLM responsible for estimating persona attributes, we conducted interviews again sequentially with the same interviewees. In this case as well, we conducted 39 interviews similarly. We instructed participants to provide the same responses under the few-shot condition as they did under the zero-shot condition. Because this paper focuses on persona attribute estimation from the dialogue history, rather than on changes in the interviewees

[1] https://openai.com/index/hello-gpt-4o/
[2] https://github.com/langchain-ai/langgraph

Table 1: Examples of Improvements in Persona Attribute Estimation Before and After Few-shot Introduction(Example 1: “School Teacher” domain)

	Utterance	Estimated Persona Attribute
System	What subjects were you particularly good at, when you were in school?	
Interviewee	I was particularly good at math when I was in school.	Favorite subjects in school
System	Could you tell me which subjects you were particularly good at in school?	
Interviewee	I was particularly good at computer science courses when I was in school.	Favorite subjects in school

(a) Excerpted Few-shot Prompt Examples Used in the Experiment(“IT engineer” domain)

Zero-shot		Few-shot	
Observed Dialogue			
What was your strongest subject in school?		What was your strongest subject in school?	
I was good at Japanese language and literature classes when I was in school.		I was good at Japanese language and literature classes when I was in school.	
Persona Attribute Estimation Results			
Human-annotated Persona Attribute (Ground Truth)	LLM Persona Attribute Estimation	Human-annotated Persona Attribute (Ground Truth)	LLM Persona Attribute Estimation
Future vision Family Promotion or job change Hobbies Past Career Favorite subjects in school	Future vision Family Hobbies Career/Workplace	Future vision Values on Education Promotion or job change Hobbies Past Career Favorite subjects in school	Future vision Values on Education Promotion or job change Hobbies Past Career Favorite subjects in school

(b) Model Estimations and Human-annotated Ground Truth (red colored persona attribute indicates the difference in persona attribute estimation between zero-shot and few-shot)

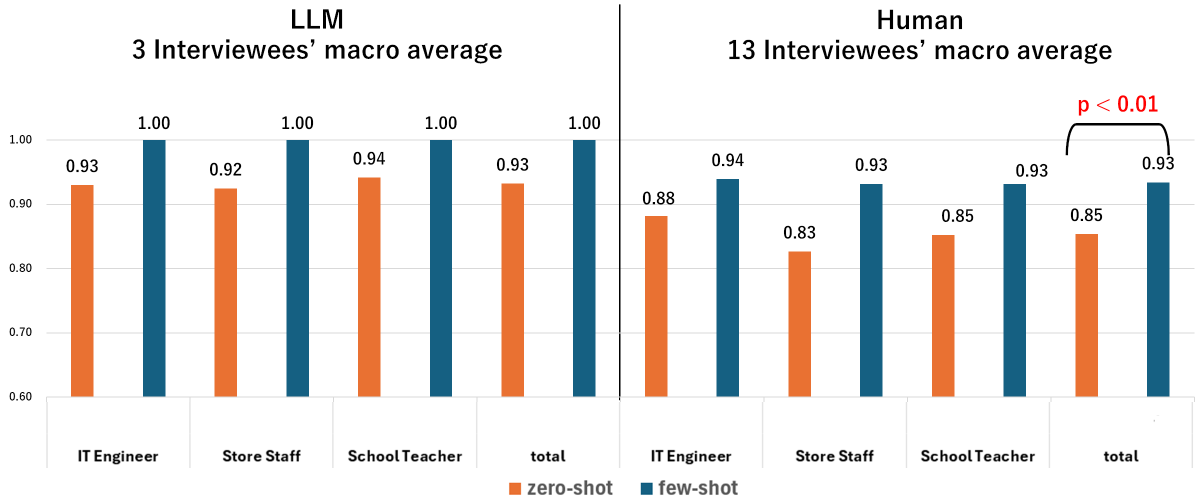

Figure 4: Evaluation Results of Estimating Persona Attributes (macro average of thirteen interviewees in each domain
)

themselves, we consider the impact of interview order on the evaluation to be small.

To avoid the examples included in few-shot prompting being identical to the interviews being evaluated, the examples were created from domains different from the evaluation target domain. Specifically, for experiments in the “IT engineer” and “store staff” domains, we used few-shot examples from the “school teacher” domain, and for experiments in the “school teacher” domain, we used few-shot examples from the “IT engineer” domain. This enabled evaluation of few-shot performance under settings where the evaluation target and few-shot domains were not identical. This design also indicates that the framework can be extended to a new domain simply by adding examples, which we consider far more scalable than traditional rule-based or domain-specific dialogue systems. The few-shot

examples were created based on preliminary experiments (3 people per domain) conducted before the main experiment. From dialogue logs obtained in the preliminary experiments, we distinguished items appropriate as persona attributes from items that should not be treated as attributes, and provided examples as positive and negative cases for persona attribute estimation. The negative examples are intended to illustrate typical responses that should not be treated as persona attributes, and they are designed to cover representative patterns of ambiguous, indirect, and non-attribute-based responses that frequently occur in actual interviews (e.g., generalizations, scene descriptions, impressions only, and abstract evaluations). Although the selection of these examples is not fully systematized, they reflect typical patterns of frequently occurring ambiguous responses.

To evaluate how accurately the proposed interview system could estimate interviewee persona attributes, we used recall, precision, and F-score as evaluation metrics. In this study, human evaluators reviewed the entire dialogue history and defined inferable persona attributes of the interviewee as the ground truth. One annotator was assigned to each domain to create the ground-truth data. Although this process involves subjective judgment, the interview questions are designed to elicit clear answers; therefore, we consider it straightforward to decide whether persona attributes can be inferred from the responses, and we regard the remaining ambiguity in ground-truth creation as extremely small. The objective of this evaluation is to measure how closely the LLM can approach such human annotation. The persona attributes output by the system after the interview conclusion were treated as estimation results. Recall was calculated with the number of persona attributes defined as ground truth as the denominator and the number of persona attributes matching the system's estimation results as the numerator. A predicted persona attribute is not judged by exact string matching. Because human answers are diverse, we treat a prediction as correct when the interview history allows the judgment that the interviewee possesses the corresponding persona attribute. Therefore, recall represents a value indicating to what extent the system could identify persona attributes that were inferable through the interview. On the other hand, precision was calculated with the total number of persona attributes estimated by the system as the denominator and the number of persona attributes matching the ground truth as the numerator. Therefore, precision is an indicator showing what proportion of the persona attributes estimated by the system were appropriate. F-score was calculated as the harmonic mean of Precision and Recall. Using these evaluation metrics, we evaluated which persona attributes the system could estimate and how accurately and comprehensively during the interview process.

Finally, based on a total of 39 interviews conducted by 39 interviewees (13 per domain), each with a distinct persona profile, [3] across three domains, we constructed a distribution of statistics under the null hypothesis to examine whether there is a difference in the F-scores between zero-shot and few-shot conditions. Specifically, we grouped the 78 F-scores obtained under both conditions into a single set, randomly reassigned them into two groups, and calculated the mean difference between the two groups. By repeating this process 1,000 times, we estimated the distribution of the mean difference that would be obtained if the null hypothesis were true. We calculated the p-value by evaluating how extreme the observed mean difference between zero-shot and few-shot was within this estimated distribution. If the resulting p-value fell below the significance level of 0.05, we conclude that it was unlikely the observed difference arose by chance.

7.2 Evaluation Results

Figure 4 presents the evaluation results in terms of macro average of F-score for the "IT engineer," "store staff," and "school teacher" domains. Figure 4 gives the evaluation results for 13 human interviewees in each domain, while for comparison purposes, it also gives the evaluation results for three LLM-based pseudo interviewees in each domain.

First, from the comparison of human interviewees and LLM-based pseudo interviewees under the zero-shot condition, a tendency was observed for persona attribute estimation to become more difficult in dialogues with human subjects. For example, in the "IT engineer" domain, the macro average F-score under the zero-shot condition was 0.93 for the LLM pseudo interviewee but decreased to 0.88 for human interviewees. Similarly, in the "store staff" domain, the LLM pseudo interviewee achieved 0.92 while the human interviewee

[3] Details on the 39 interviewees are presented in section B of Appendix.

achieved 0.83. These results suggest that estimating persona attributes becomes more difficult in interviews with human subjects.

Next, for human interviewees, the few-shot condition surpassed the zero-shot condition in all domains. In fact, the p-value was under 0.01, indicating that there is a significant difference in F-scores between zero-shot and few-shot. The macro average F-score consistently improved across all domains for human interviewees, suggesting that introducing few-shot prompting contributes to improved persona attribute estimation performance for human utterances. On the other hand, for LLM-based pseudo interviewees, while improvement in macro average F-score was observed with the few-shot condition in the store staff domain, slight decreases in macro average F-score were confirmed in the "IT engineer" and "school teacher" domains.

Table 1 shows a case comparing persona attribute estimation results between zero-shot and few-shot conditions in the "school teacher" domain. In this case, under the few-shot condition, dialogue examples belonging to the "IT engineer" domain are presented as prompts in advance. In these few-shot examples, the correspondence relationship between the interviewer's questions and persona attributes extracted from the corresponding utterance content is explicitly shown. In the conversation under the zero-shot condition, although the interviewee responds as "I was good at Japanese language and literature classes when I was in school.", the LLM could not identify the persona attribute corresponding to "Favorite subjects in school" (in red colored character) from this utterance. On the other hand, under the few-shot condition, a similar utterance about favorite school subjects was obtained, and in this case, it can be seen that the LLM correctly estimated the corresponding persona attribute "Favorite subjects in school" (in red colored character).

This result suggests that the few-shot condition enabled appropriate attribute estimation by providing the LLM in advance with decision criteria for mapping utterances to corresponding persona attributes.

Table 4 and Table 5 in Section A of Appendix show that, in the "store staff" and "IT engineer" domains, few-shot prompts enable the LLM to more consistently associate the utterances with the appropriate persona attribute which is not reliably identified under the zero-shot condition. In Table 4 and Table 5, again, red colored persona attributes "Workplace relationships" in Table 4 and "Basic Personal Information" in Table 5 respectively are among those correctly estimated only with the few-shot condition but not with the zero-shot condition.

8 Conclusion

In this study, we implemented an LLM interviewer based on semi-structured interviews and evaluated persona attribute estimation from dialogue history and slot information, for both LLM-based pseudo interviewees and human interviewees, with and without few-shot prompting. Estimation was more difficult with human interviewees. Few-shot prompting improved performance for humans in all domains, mainly in recall, but had mixed effects for LLM-based pseudo interviewees. Case analysis confirmed that examples can improve estimation.

As future work, as discussed in section D of Appendix, in order to make the semi-structured interview a more practical one, we will introduce more structured persona design. By this, it becomes easier for the interviewer LLM to efficiently organize the questions in the dialogue, where, out of those 10 initially set persona attributes, it is expected that the number of unexplored ones will decrease to almost zero.

Limitations

In this paper, interviews were conducted with 13 interviewees for each of the three domains (39 interviewees in total). Moving forward, it will be necessary to increase the number of interviewees and conduct experiments on a larger scale. A potential risk of this work is the impact on system stability resulting from the domain change. While we are currently conducting experiments using three domains, we should conduct further experiments with additional domains to account for a wider range of interviewees.

References

Pooja Rao S B, Manish Agnihotri, and Dinesh Babu Jayagopi. 2020. Automatic follow-up question generation for asynchronous interviews. In *Proc. IntelLanG*, pages 10–20.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are

few-shot learners. In *Proc. 33rd NeurIPS*, pages 1877–1901.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proc. 58th ACL*, pages 107–121.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and 3 others. 2014. Simsensei kiosk: a virtual human interviewer for healthcare decision support. In *Proc. 13th AAMAS*, page 1061–1068.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards LLM-driven dialogue state tracking. In *Proc. EMNLP*, pages 739–755.

NG Fielding, editor. 2003a. *Interviewing*, volume I. Sage Publications.

NG Fielding, editor. 2003b. *Interviewing*, volume II. Sage Publications.

Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. 2023. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In *Proc. 37th PACLIC*, pages 113–124.

Friedrich Geiecke and Xavier Jaravel. 2024. Conversations at scale: Robust ai-led interviews with a simple open-source platform. *Social Science Research Network*, pages 1–73.

Ryo Hasegawa, Yijie Hua, Takehito Utsuro, Ekai Hashimoto, Mikio Nakano, and Shun Shiramatsu. 2025. A dialogue system for semi-structured interviews by LLMs and its evaluation on persona information collection. In *Proc. IWSDS 2025*, pages 39–59.

Ekai Hashimoto, Mikio Nakano, Takayoshi Sakurai, Shun Shiramatsu, Toshitake Komazaki, and Shiho Tsuchiya. 2025. A career interview dialogue system using large language model-based dynamic slot generation. In *Proc. 31st COLING*, pages 1562–1584.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proc. 61st ACL*, pages 936–950.

Jiaxiong Hu, Jingya Guo, Ningjing Tang, Xiaojuan Ma, Yuan Yao, Changyuan Yang, and Yingqing Xu. 2024. Designing the conversational agent: Asking follow-up questions for information elicitation. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proc. 24th SIGDIAL*, pages 216–228.

Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job interviewer android with elaborate follow-up question generation. In *Proc. ICMI*, page 324–332.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proc. 23rd SIGDIAL*, pages 336–350.

Michael Johnston, Patrick Ehlen, Frederick G. Conrad, Michael F. Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. 2013. Spoken dialog systems for automated survey interviewing. In *Proc. 14th SIGDIAL*, pages 329–333.

Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. 2016. Small talk improves user impressions of interview dialogue systems. In *Proc. 17th SIGDIAL*, pages 370–380.

Keigo Komada, Kaori Abe, Shoji Moriya, and Jun Suzuki. 2024. Dynamic slot-making-and-filling method for improving long-term dialogue consistency. In *Proc. JSAI*, page JSAI2024(0):4Xin2102–4Xin2102 (in Japanese).

Fuminori Nagasawa, Shogo Okada, Takuya Ishihara, and Katsumi Nitta. 2024. Adaptive interview strategy based on interviewees' speaking willingness recognition for interview robots. *IEEE Transactions on Affective Computing*, 15(3):942–957.

Angelina Parfenova. 2024. Automating the information extraction from semi-structured interview transcripts. In *Proc. WWW*, page 983–986.

Jost Schatzmann and Steve Young. 2009. The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):733–747.

A.B. Siddique, Fuad Jamour, and Vagelis Hristidis. 2021. Linguistically-enriched and context-awarezero-shot slot filling. In *Proc. WWW*, page 3279–3290.

Ming-Hsiang Su, Chung-Hsien Wu, and Yi Chang. 2019. Follow-up question generation using neural tensor network-based domain ontology population in an interview coaching system. In *Interspeech*, pages 4185–4189.

Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, and Huai-Hung Huang. 2018. Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching. In *Interspeech*, pages 1006–1010.

Guangzhi Sun, Shutong Feng, Dongcheng Jiang, Chao Zhang, Milica Gasic, and Phil Woodland. 2024. Speech-based slot filling using large language models. In *Findings of ACL*, pages 6351–6362.

Nicolas Wagner and Stefan Ultes. 2024. On the controllability of large language models for dialogue interaction. In *Proc. 25th SIGDIAL*, pages 216–221.

Tom Wengraf. 2001. *Qualitative Research Interviewing*. Sage Publications.

Jie Zeng, Yukiko Nakano, and Tatsuya Sakato. 2023. Question generation to elicit users' food preferences by considering the semantic content. In *Proc. 24th SIGDIAL*, pages 190–196.

Yuki Zenimoto, Mariko Yoshida, Ryo Hori, Mayu Urata, Aiko Inoue, Takahiro Hayashi, and Ryuichiro Higashinaka. 2025. Automated administration of questionnaires during casual conversation using question-guiding dialogue system. In *Proc. SEMDIAL*, pages 115–124.

A Additional Figures and Tables

Figure 5 contrasts typical response characteristics of human interviewees and an LLM-based pseudo interviewee in the semi-structured interview setting. Table 2 provides example dialogue excerpts comparing responses from a human interviewee and an LLM-based pseudo interviewee. As a complement for Figure 4 presenting F-score only but not recall nor precision, Figure 6 presents the evaluation results in terms of macro average of recall / precision / F-score over 13 human interviewees in each of "IT engineer," "store staff," and "school teacher" domains. As shown in the figure, in all the domains, few-shot achieved improvement over zero-shot in F-scores, where their improvements in recall are more dominant than those in precision in each of the three domains. Table 3 summarizes an example predefined persona used in the experiments, listing persona attributes and their values for an IT engineer ("Shota Suzuki"). Table 4 and Table 5 present additional case studies from the store staff and IT engineer domains, showing how few-shot examples help map utterances to the persona attributes "Workplace relationships" and "Basic Personal Information," respectively.

B Details on Participants

39 human participants took part in the experiments described in this paper (13 participants in each of the three domains). Each participant was assigned a distinct persona profile and did not participate in more than one domain. All participants were students and received no financial compensation. Prior to the experiments, we explained the purpose and importance of this study, and all participants agreed to take part in the evaluation on a voluntary basis. The following instructions were given to the participants before the interviews. Participants were asked (i) to respond to the system's questions as interviewees, (ii) to answer by role-playing the pre-defined persona assigned to them, and (iii) not to disclose any personal information beyond what was specified in the persona description.

C Few-shots versus CoT Prompts

What we have given to LLMs as few-shots are far beyond conventional few-shots but they are nearly equivalent to CoT studied in Zenimoto et al. (2025) both in their quantity and quality. Our few-shots consist of cross domain dialogue history of three interview sessions as well as all the pairs of slot names and slot values automatically extracted by a zero-shot LLM from the given dialogue history. Here, each interview session corresponds to an interview with an interviewee. All the pairs of slot names and slot values extracted from the given dialogue history, on the other hand, are divided into positive ones corresponding to persona attributes of each interviewee and the remaining negative ones not corresponding to persona attributes. (For example, few-shots collected from the school teachers domain consist of 31 persona attributes and 43 of those other than persona attributes, while those collected from the store staff domain consist of 31 persona attributes and 44 of those other than persona attributes.) Those plenty of few-shots are valuable information source that represent criteria when judging whether a pair of a slot name and a slot value is appropriate as a persona attribute, which can be regarded as a CoT.

D Persona Attributes Employed in Experiments

The 10 persona attributes are manually set in advance before we start interview experiments. The subset of those 10 persona attributes manually set in advance and appearing in the actual dialogue are only a part of the whole ground truth persona attributes. The whole ground truth persona attributes include the subset of those 10 and also the additional ones that are manually selected through careful examination of the actual dialogue history.

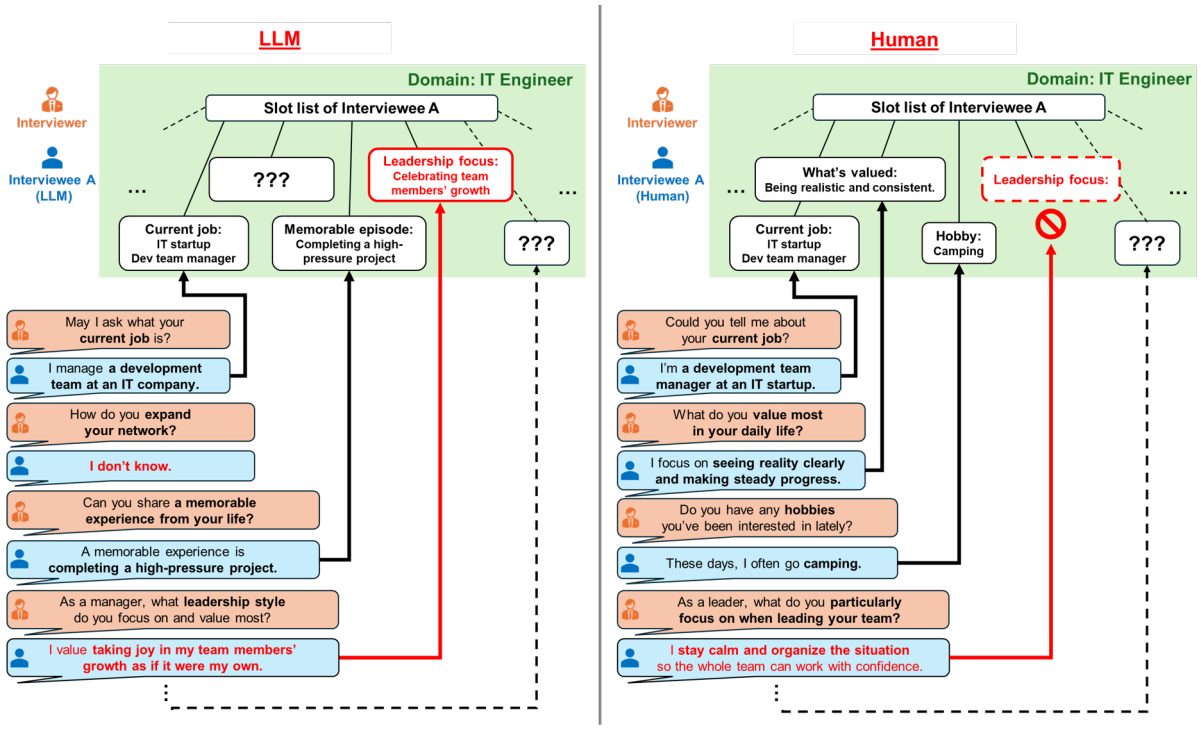

Figure 5: Diagram of Semi-Structured Interviews with Human and LLM Interviewees by Zero-shot Prompts

Speaker	Utterance	Estimated Persona Attribute
System:	As a manager of the development team, what leadership style do you value most?	
Interviewee:	I value a style that is attentive to others and takes genuine delight in the growth of team members as if it were my own.	[[Correct]]: Leadership Style

(c) Dialogue between the System and the LLM Interviewee

Speaker	Utterance	Estimated Persona Attribute
System:	What do you particularly focus on or keep in mind when leading your team as a leader?	
Interviewee:	I will maintain a level head and organize the situation so the entire team can move forward with confidence.	[[Incorrect]]: "None"

(d) Dialogue between the System and the Human Interviewee

Table 2: The differences of responses between Human-interviewee and LLM-interviewee

(shown in Table 1, Table 4, and Table 5).

As clearly shown in those tables, regarding the comparison between zero-shot and few-shot, we are successful in achieving higher recall and hence higher F-score in the task of persona estimation through semi-structured interview. However, in order to make the semi-structured interview a more practical one, we will introduce more structured persona design. By this, it becomes easier for the interviewer LLM to efficiently organize the questions in the dialogue, where, out of those 10 initially set persona attributes, it is expected that the number of unexplored ones will decrease to almost zero.

E Topic Selection Strategy in Dialogues

In the dialogue session of the proposed method, the following topic selection strategy is employed.

(1) First, it is randomly selected whether the follow-up question strategy is to be selected or not.

(2a) when the follow-up question strategy is selected, a new slot that follows the current thread of the ongoing dialogue is generated by the slot generator LLM, and then a new question for filling the newly generated slot is composed by the question composer LLM.

(2b) even when the follow-up question strategy is

ID	Persona Attributes (given in advance)	Values of "Shota Suzuki" (given in advance)
1	Basic Personal Information (Name, Age, Hometown, Gender)	Shota Suzuki, 25 years old, Tokyo, Male
2	Personality	Strong in logical thinking and highly intellectually curious. Once focused, tends to become deeply absorbed in tasks and persistently works through problems until they are solved.
3	Past Career	After graduating from a university with a major in information science, joined a current major tech venture as a new graduate and is now in the third year. In the university laboratory, engaged in research on image recognition technology and presented findings at academic conferences.
4	Current Career	Working as a backend engineer at a company that operates one of Japan's largest social networking services. Mainly responsible for new API development and maintenance/operation of microservices.
5	Vision for the current career	Wants to be involved in designing the overall architecture of large-scale services. Aims to become a tech lead who can participate from the stage of technology selection.
6	Concerns and Dissatisfaction	Because the organization is large and vertically divided, it's difficult to see how one's own work contributes to the overall product. Feels stressed by the large number of coordination tasks such as meetings and documentation, beyond writing code.
7	Thoughts on Promotion or Career Change	Has a professional orientation toward pursuing technical expertise rather than management. In the future, is also considering working at an early-stage startup with greater autonomy.
8	Hobbies and Personal Life	Hobbies include intellectually challenging board games and weekend bouldering. Often spends days off reading trending technical books or discussing interesting technologies with friends.
9	Family	Currently living alone near the company, having moved out from the family home in Tokyo. Family includes parents and a younger sister who is a university student.
10	Memorable Episodes	In his first year at the company, a feature he developed was released, and he saw users on social media saying things like, "This feature is really useful." It made him realize that the code he wrote had a real impact on people's lives, and it gave him a strong sense of purpose and fulfillment.

Table 3: List of persona attributes for interviewees and values for the first interviewee, IT engineer "Shota Suzuki"

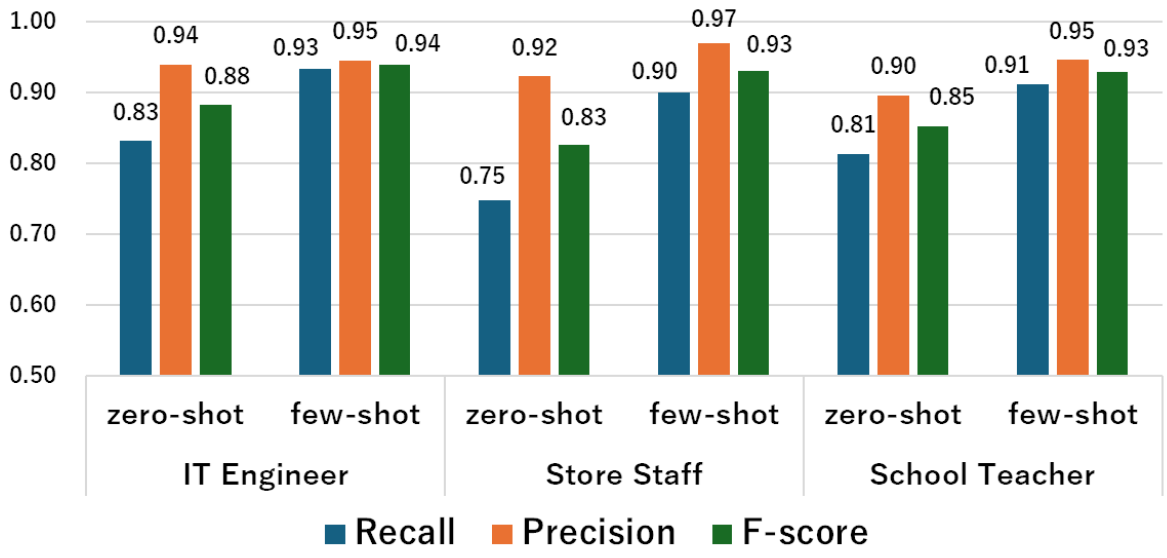

Figure 6: Evaluation Results of Estimating Persona Attributes (macro average of Recall / Precision / F-scores over interviewees)

NOT selected, among the pool of topics for shifting from current topic to another one, considering the history of topics already covered in this dialogue so far, one remaining topic is randomly selected. A new slot corresponding to the selected topic is then generated by the slot generator LLM, and a question for filling that slot is composed by the question composer LLM.

Table 4: Examples of Improvements in Persona Attribute Estimation Before and After Few-shot Introduction(Example 2: “Store Staff” domain)

	Utterance	Estimated Persona Attribute
System	Could you share any particularly memorable experiences or thoughts you have regarding interpersonal relationships in the educational setting?	
Interviewee	One of my former students, now an adult, went out of their way to come and say hello, telling me, “You were scary, but I am who I am today because you taught me so strictly back then.	Workplace Relationships
System	How do you feel about your relationships with colleagues and supervisors in the educational setting?	
Interviewee	I don’t really have any particular thoughts about my relationship with my boss.	None

(a) Excerpted Few-shot Prompt Examples Used in the Experiment(“school teacher” domain)

Zero-shot		Few-shot	
Observed Dialogue			
How do you feel about relationships in the workplace? For example, please tell us about any memorable experiences you’ve had with colleagues or supervisors.		How do you feel about relationships in the workplace? For example, I’d love to hear about your relationships with colleagues and supervisors.	
We have many regular customers, and I enjoy working in this friendly, relaxed atmosphere.		We have many regular customers, and I enjoy working in this friendly, relaxed atmosphere.	
Persona Attribute Estimation Results			
Human-annotated Persona Attribute (Ground Truth)	LLM Persona Attribute Estimation	Human-annotated Persona Attribute (Ground Truth)	LLM Persona Attribute Estimation
Current career Workplace relationships Personality Past Career Basic Personal Information	Career/Workplace Personality Basic Personal Information	Current career Workplace relationships Personality Past Career Basic Personal Information	Current career Workplace relationships Personality Past Career Basic Personal Information

(b) Model Estimations and Human-annotated Ground Truth(red colored persona attribute indicates the difference in persona attribute estimation between zero-shot and few-shot)

Table 5: Examples of Improvements in Persona Attribute Estimation Before and After Few-shot Introduction(Example 3: “IT engineer” domain)

	Utterance	Estimated Persona Attribute
System	Could you tell me what kind of neighborhood you live in?	
Interviewee	I live in Kanagawa Prefecture	Basic Personal Information
System	Could you tell me about where you’re from and the environment you grew up in?	
Interviewee	I’m from Chiba Prefecture	Basic Personal Information

(a) Excerpted Few-shot Prompt Examples Used in the Experiment(“school teacher” domain)

Zero-shot		Few-shot	
Observed Dialogue			
First, I’d like to ask you a basic question: Where do you live?		Could you start by briefly introducing yourself?	
I’m from Saitama Prefecture		My name is Takuya Nakamura, I’m a 38-year-old man. I’m from Saitama Prefecture.	
Persona Attribute Estimation Results			
Human-annotated Persona Attribute (Ground Truth)	LLM Persona Attribute Estimation	Human-annotated Persona Attribute (Ground Truth)	LLM Persona Attribute Estimation
Personality Side job Basic Personal Information Promotion or job change Hobbies Family	Personality Career/Workplace Hobbies Family	Personality Decision-making process Basic Personal Information Promotion or job change Hobbies Family	Personality Decision-making process Basic Personal Information Promotion or job change Hobbies Family

(b) Model Estimations and Human-annotated Ground Truth (red colored persona attribute indicates the difference in persona attribute estimation between zero-shot and few-shot)