# LM-Interview: An Easy-to-use Smart Interviewer System via Knowledge-guided Language Model Exploitation

**Hanming Li[1], Jifan Yu[2], Ruimiao Li[2], Zhanxin Hao[2],**
**Xuan Yan[2], Jiaxin Yuan[2], Bin Xu[1], Juanzi Li[1], Zhiyuan Liu[1]**
[1]Department of Computer Science and Technology, BNRist, Tsinghua Univeristy
[2]Tsinghua University
{lhm22, yujf21, lrm20, zhanxin_hao, yan-x21, yuanjx21}@mails.tsinghua.edu.cn
{xubin, lijuanzi, liuzy}@tsinghua.edu.cn

## Abstract

Semi-structured interviews are a crucial method of data acquisition in qualitative research. Typically controlled by the interviewer, the process progresses through a question-and-answer format, aimed at eliciting information from the interviewee. However, interviews are highly time-consuming and demand considerable experience of the interviewers, which greatly limits the efficiency and feasibility of data collection. Therefore, we introduce LM-Interview[1], a novel system designed to automate the process of preparing, conducting and analyzing semi-structured interviews. Experimental results demonstrate that LM-Interview achieves performance comparable to that of skilled human interviewers.

## 1 Introduction

Interviews are a widely employed method that exerts a profound influence in the field of qualitative research. The central concept of structured interviews is to ensure that each interview is conducted with exactly the same questions presented in the same order. This standardization ensures that answers can be reliably aggregated and that comparisons can be confidently made between different subgroups within the sample or across various survey periods. On the basis of structured interviews, **semi-structured interviews** take a step further by breaking the constraints of a fixed set of questions and predefined order, posing probing questions to the details emerged during the interviews, therefore enabling the uncovering of deeper knowledge and more profound associations while maintaining a similar level of comparability between samples as structured interviews. However, conducting semi-structured interviews necessitates extensive involvement of experienced researchers, which severely limits the efficiency of data collection, hence the generalizability of the researches.

For a seemingly viable solution to automate the process, the Task-Oriented Dialogue (TOD) system (Wen et al., 2016; Kwan et al., 2023; Hosseini-Asl et al., 2020) aims to respond to user inputs within a predefined action space. By parsing natural language utterances into specific ontology, the system then tracks the state and selects an action to generate a response that fulfills the expected functions. However, applying such a pipeline is not entirely satisfactory, due to the challenging nature of semi-structured interviews as follows:

**(1) Control by Interviewers.** TODs are specifically designed to facilitate user-initiated tasks. In contrast, interviewees in semi-structured interviews usually lacks a specific agenda, necessitating that the system exert control over the interview process, which should be guided by a comprehensive, pre-established plan.

**(2) Flexibility of Actions.** While the utterances of interviewers can generally be categorized into actions such as responding or posing probing questions, these actions tend to be more *experiential* rather than *factual*. That is to say, the boundaries and expected behaviors are not strictly defined, which complicates the definition of the action space when implementing a system.

**(3) Necessity of Analysis.** To effectively support arguments or yield insights, the data collected must first undergo thorough analysis, which is often overlooked in previous dialogue systems primarily focusing on the mere exchange of information.

**Presented system.** In this paper, we introduce LM-Interview, a system designed to support qualitative researchers throughout the procedure of semi-structured interviews. By leveraging knowledge-guided language model exploitation, LM-Interview addresses each of the three identified gaps through strategically designed modules, the workflow of which aligns with the typical process division for conducting semi-structured interviews described in classical literature (Kvale, 2012). Qualitative

[1]https://github.com/HwHunter/LM-Interviewer



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*
*System Demonstrations*, pages 520–528

researchers can utilize our system to construct the interview guides before interview, then gather extensive data by LLM-driven interviews without the need for human labor, and finally, gain insights from the system's analysis of the interview data to advance their researches.

**Contributions.** (1) We propose the use of a knowledge-guided language model to automate the process of conducting semi-structured interviews. (2) We implement LM-Interview, a comprehensive system designed to supporting qualitative researchers throughout the entire process of designing, conducting, and analysing interviews. (3) We conduct experiments demonstrating that the system achieves a level of performance comparable to that of experienced human interviewers.

## 2 The Interview System

The typical process (Kvale, 2012) of carrying out an interview is dividing it into three stages: (1) Constructing the **Interview Guide** before the interview, (2) Chatting with the interviewee to **Gather Information** during the interview, and (3) After the interview, encoding the discourse and conduct **Conversation Analysis**. Following this widely-applied paradigm, we design multiple modules for all the three stages as shown in Figure 1, which are all empowered by language model coordination.

### 2.1 Pre-Stage: Guide Construction Module

Although a competent interviewer adapts to the actual course of semi-structured interviews, adjustments must still be made within or at least around a predefined question framework, which is called **interview guide** (Naz et al., 2022; Williams, 1988). Predictably, the interview guide plays a crucial role in semi-structured interviews, which is why several authoritative sources recommend memorizing it prior to conducting the interviews (Lareau, 2021; Kvale, 2012).

A well-designed one should contain open-ended questions organized in two layers: (1) the **main questions**, which address the broad topics of interest to guide the overall direction of the conversation, and are provided by the researchers when using our system; and (2) the follow-up questions, or **probes**, which arise from main questions and are design to delve deeper into specific points that emerge as particularly valuable during the discussion, which are generated with this Guide Construction Module.

Formally speaking, given a list of main questions $\{M_i\}$, the Guide Construction Module generates multiple probes $\{P_i\}$ for each $M_i$, that is

$$\texttt{GCM} : M_i \rightarrow \{P_{i,j}\}_{j=1}^{n_i} \tag{1}$$

to form a complete interview guide:

$$\texttt{Guide} = \bigcup_i (\{M_i\} \cup \{P_{i,j}\}_{j=1}^{n_i}) \tag{2}$$

Such generating involves addressing two gaps between the two layers of questions. **(1) General vs. Specific:** main questions establish the framework of the interview, while probes must delve into the details of each main question, necessitating a thorough understanding of them; **(2) Anticipated vs. Actual**: the main questions outline the expected interview issue, while probes must cover potential valuable points that emerge during the interview, requiring prediction to the actual process. Following Chain-of-Thought prompting (Wei et al., 2022), we develop a step-by-step approach to generate the interview guide from main questions provided by the researchers, which is illustrated in Figure 2. Specifically, in a multi-turn dialogue with the agent, we instruct it to **(1) Main Questions Comprehension**, which address the first gap, then **(2) Potential Direction Prediction** of the interview, which address the second gap, and finally **(3) Probes Generation** for each main question. We also design an extra step, **(4) Quantitative Metrics Configuration**, for organizing analysis of the interview, which will be discussed later in 2.3.

### 2.2 Major-Stage: Dialogue Module

Structured interviews have the primary benefit that they allow interviews to focus on the planned route, while still giving the interviewer the autonomy to explore relevant ideas that emerge during the interview (Adeoye-Olatunde and Olenik, 2021). However, such merits also lay challenges for even experienced human interviews of controlling the *tempo*, i.e. the balance between two conflicting aspects (1) adhering the pre-made guide and (2) probing emerged details for additional information.

Such requirements require delicate control over the behaviors of interview agent. Following the famous *state, action, reward* paradigm of reinforce learning (Kaelbling et al., 1996), the dialogue during an interview is formed as a multi-turn conversation within the *context, action, information* process:

**Context** consists of alternating utterances between the interviewee and the agent interviewer.

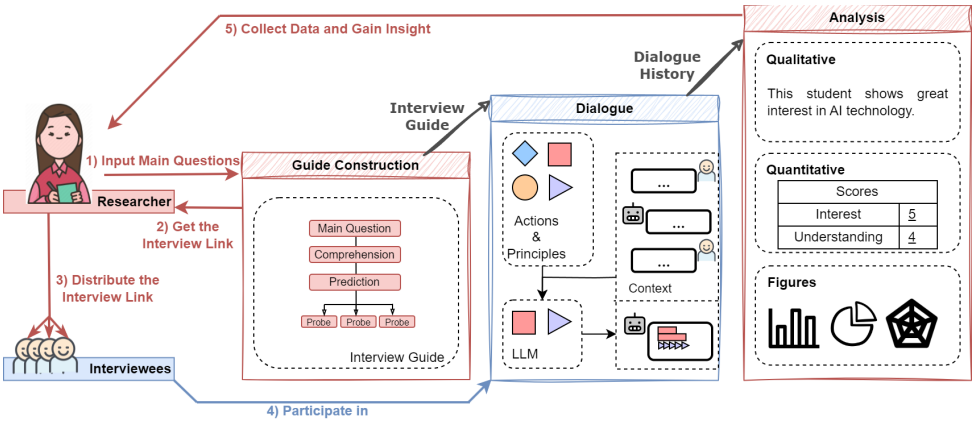


Figure 1: Workflow of LM-Interviewer.

For the $i$-th turn, denoting the question asked by the agent interviewer as $Q_i$, the answer by the interviewee $A_i$, which can be formed as

$$\texttt{Context}_i = \{Q_1, A_1, ..., Q_{i-1}, A_{i-1}\} \quad (3)$$

**Actions**, given $Context_i$, are the behaviors included in the agent's next question $Q_i$. By defining types of actions and the conditions under which each action is applicable, we can finely tune the agent's behavior, thus to maximize expected collected **Information**. For formally representation, given the $Context_i$, the agent will pose a question $Q_i$, which sequentially includes multiple actions

$$Q_i = \{\texttt{Action}_{i,1}, ..., \texttt{Action}_{i,n_i}\} \quad (4)$$

and the actions are chosen under policy $\mathcal{P}$

$$\mathcal{P} : \texttt{Context}_i \rightarrow \{\texttt{Action}_{i,j}\} \quad (5)$$

Given the definition above, adjusting the behaviors of the agent involves defining the action space and establishing the policy. For action space, to fulfill the two conflicting aspects of a semi-structured interview both, we define two actions for each, which are briefly summarized in Table 1, and illustrated with an example in Figure 3, while the policy is encoded in the prompt in the form of *principles*, which specify the behaviors and applicable conditions through a set of natural language guidelines summarized from (Lareau, 2021) by human experts for each type of action. The complete list of principles can be found in Table 2 in appendix.

### 2.3 Post-Stage: Analysis Module

The raw output from the dialogue module consists of a series of questions and answers, which cannot be leveraged without analysis (Lillis, 1999; Rabiee, 2004; Roulston, 2011) in various interview application scenarios. For example, in qualitative research, interviewers should write "analytic memos" regularly during the data collection process (Lareau, 2021). We implement the analysis module from both qualitative and quantitative dimensions.

| | |
|---|---|
| **Focus on the Plan** | |
| Querying | Pose a question by the guide |
| Advancing | Introduce the next topic |
| **Probe for Details** | |
| Probing | Ask about emerged details |
| Responding | React and respond actively |

Table 1: The action space of the dialogue module. The actions are categories by the two aspects of semi-structured interviews, along with brief descriptions.

**Qualitative dimension.** The system can automatically summarize the conversational information (Ma et al., 2022). Similar to analytic memo, the summary contains the key elements and discoveries about the interview.

**Quantitative dimension.** The experiment results are hard to analyze qualitatively as they scale up, which usually leads to loss of generalizability (Holton and Burnett, 2005). Threrefore, we use LLMs to analyze the interviews and obtain numerical data, or **scores**, on the metrics proposed in the last stage of guide construction. **Explanations** for the scores will be generated along with them to enhance the credibility.

Given that requirements of analysis differ across various applying scenarios of interviews, such multi-dimensional implementation grants our system enhanced adaptability. Data collection is heavily based on interviews and both qualitative and quantitative dimensions can offer valuable insight

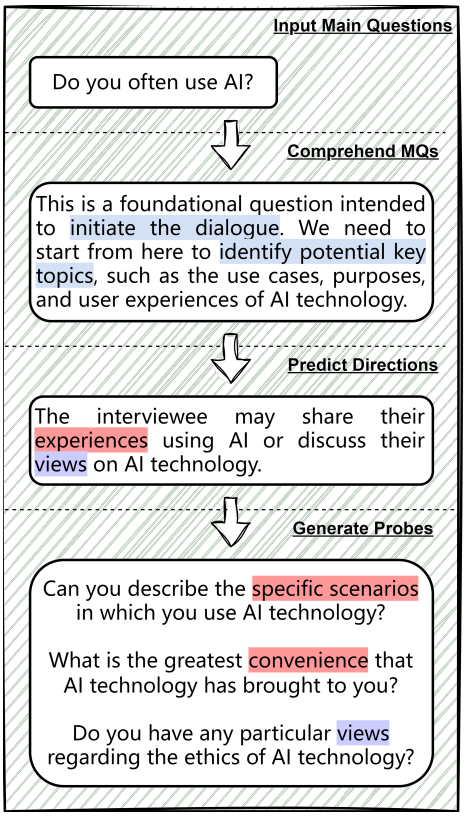


Figure 2: An illustration of constructing the interview guide, which is the combination of (1) and (4) by definition. Key points in (2) and relating information between the (3) and (4) are highlighted.

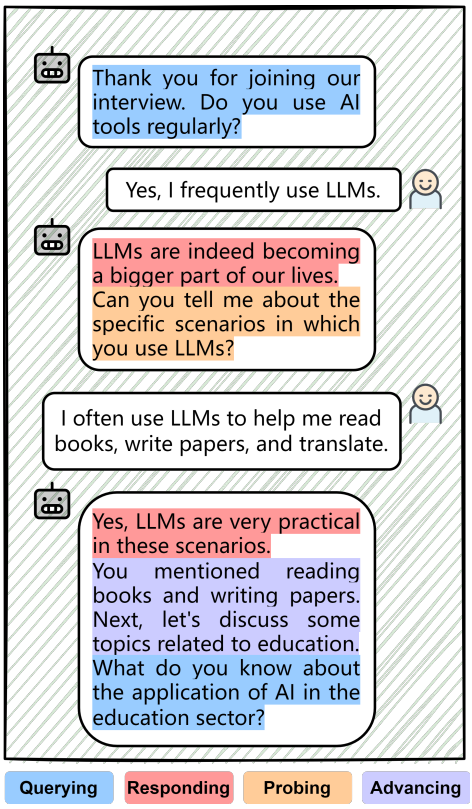


Figure 3: Actions in questions posed by the interview agent, which are highlighted with different colors.

into dialogue data. In scenarios where statistics itself matters, the quantitative dimension becomes particularly useful as probably a superior alternative in a certain perspective to traditional methods, e.g. scales or questionnaires, since the scores are supported by conversational information that might not be accessible through other means (Blaxter et al., 2010).

```python
@descriptive_analysis
def radar_chart(metrics: list[str]
                scores: list[list[int]]) -> str:
    # implementation of drawing a radar chart
    return path_to_chart
```

Figure 4: The reserved decorator and exemplary function signature for descriptive analysis functions. The image returned will be included in the output of analysis module.

**Descriptive Analysis.** Both qualitative and quantitative dimensions provide insights into a single individual. To depict the collective characteristics of all interviewees, we implement descriptive analysis using charts, in a *hot-swappable* manner. Specifically, in our implementation, all functions with a reserved decorator are viewed as a **descriptive analysis function**, which return the path to the chart it plots. The usage and exemplary function signature is illustrated in Figure 5. All charts from the descriptive analysis function will be presented in the final analysis result. Thus, researchers of different fields can integrate data analysis and visualization methods of their own field in our system.

## 3 Experiments

In this section, we conduct real-scenario experiments to evaluate the proposed system. Specifically, we assess the system's ability to (1) conduct and (2) analyze interviews.

### 3.1 Experimental Setup

**Interviews Setup.** The interviews were conducted in a real-world setting to evaluate the user experience of students who participated in a AI-assisted classroom (Zhang et al., 2024). We designed an interview guide and used the 17 main questions it contained as inputs for guide construction mod-

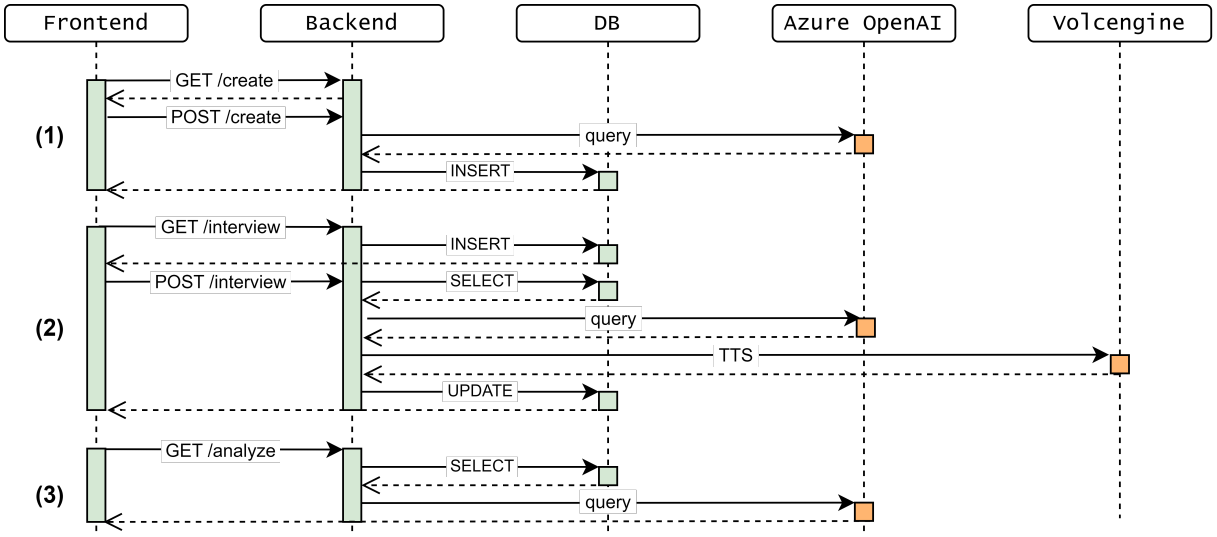


Figure 5: UML diagram for our system implementation.

ule. We recruit 7 students who are first interviewed by experienced human interviewers and then by the system one week later to avoid interference between.

**System Implementation.** The demo web application, illustrated in Figure 5, is implemented with Flask framework[2]. For the backend, we implement multiple endpoints for each modules. We deploy a sqlite[3] database to store all the data (e.g. generated interview guide, dialogue history). Only the primary key of each interview is stored in session, with which the data is retrieved from the database in each round of dialogue. For the frontend, the pages are written in HTML/CSS, communicating with the backend with HTTP requests and socket[4] for audio data. We use gpt-4-1106 with default parameters (`n = 1`, `temperature = 1.0`, `max_token = 4096`) from Azure OpenAI Service[5] as the backbone of agent without further tuning. To enhance the sense of presence, we implement ASR (Automatic Speech Recognition) and TTS (Text To Speech) during the interview process using volcengine[6].

## 3.2 Capability to Conduct Interviews

Since our system has two groups of users: researchers who design and conduct studies, and interviewees who are recruited and participate in the interviews, we evaluate the capability of our system to conduct interviews, i.e. to collect information via conversation, from two perspectives. (1) From the perspective of researchers, we analyze the ratings given by two qualitative research expert, who compared the processes of interviews conducted by humans and the system. (2) From the perspective of interviewees, we analyzed the ratings given by the interviewees in questionnaires, which are filled out after experiencing both the human and system interviews.

### 3.2.1 Evaluation Scheme

Based on theories and methods from several key texts (Willgens et al., 2016; Agostinho, 2005; Tracy, 2010; Corbin and Strauss, 2014), we have developed two sets of evaluation schemes from the perspectives of researchers and interviewees, respectively.

**Structure.** Both schemes are hierarchical, consisting of two levels of indicators. The lower-level *sub-indicators* focus on concrete technical details of the interview, allowing experts and users to evaluate more precisely. These sub-indicators are grouped and the average within each group forms the upper-level *aggregate indicators*, which are summaries of the system performance in several key aspects, making it easier to understand and analyze.

**Aggregate indicators.** As main aspects of the performance of the interviews, the same set of aggregate indicators are shared between two schemes,

[2]https://flask.palletsprojects.com/en/3.0.x/
[3]https://www.sqlite.org/
[4]https://flask-socketio.readthedocs.io/en/latest/
[5]https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models
[6]https://www.volcengine.com/

which are *Accuracy*, *Answerability*, *Organization*, *Engagement*, *Probing*.

**Sub-indicators.** Considering the different levels of knowledge and perspectives of researchers and interviewees, we designed different sets of sub-indicators for them. As for the design principle, sub-indicators for researchers are more detailed and require greater expertise on interviews, whereas those for users focus more on the experiences.

For the process of scoring sub-indicators by experts and interviewees, we adopt a five-point Likert scale as the measurement. In such scale, the values range from *1* to 5, where 3 indicates a level comparable to human performance, and higher values indicate a clearer advantage of the system. As the average of a group of sub-indicators, each aggregate indicators pertains the same constrains and meaning.

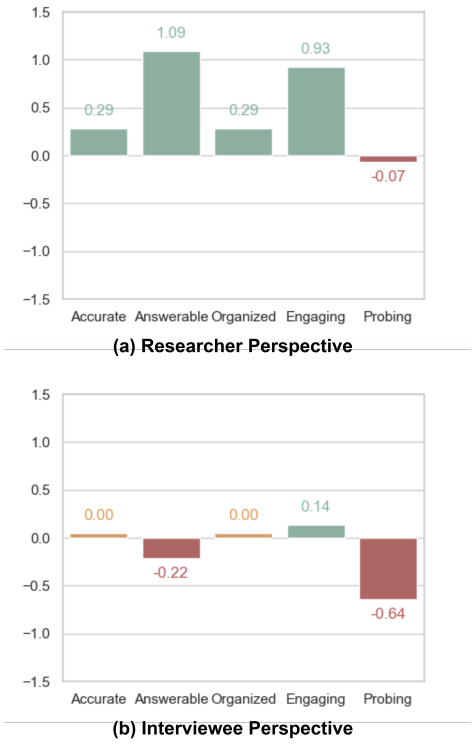


Figure 6: Ratings from both perspectives. The ratings above are shifted by -3, which means zero corresponds to "comparable to human" in the five-point scale.

#### 3.2.2 Analysis

The visualization results are shown in Figure 6, note that the scores from the two perspectives are not comparable due to the different set of sub-indicators. From the results we can acquire two observations: (1) From both the perspectives of researchers and interviewees, the system have reached a comparable overall performance to the experienced human interviewers; (2) Although the system's ability of probing is adequate, it remains its greatest weakness, which confirms our earlier point that managing the tempo is one of the biggest challenges in conducting interviews.

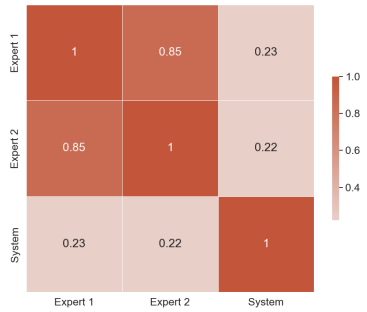


Figure 7: The heatmap of Spearman's rank correlation coefficients between the quantitative ratings given by two Experts and the system.

### 3.3 Accuracy of Quantitative Analysis

In this experiment, we evaluate the system's capacity of analysing interviews by assessing the quantitative analysis produced by the analysis module. Specifically, on the 13 metrics proposed in the guide construction module, e.g. the intensity of the interviewees' motivation to participate in the AI classroom, we calculate the Spearman's rank correlation coefficients (Spearman, 1961) between the ratings from our system and those from the human experts, which is visualized in Figure 7.

**Analysis.** The results (`corr = 0.228, p = 0.030` for Expert 1 and `corr = 0.222, p = 0.034` for Expert 2) indicate a significant weak positive correlation between the ratings given by the system and the experts, suggesting that the system has the preliminary capability to extract quantitative information from interviews.

## 4 Conclusion

We introduced LM-Interviewer, a system powered by knowledge-guided language model for automating the complete process of semi-structured interviews. With LM-Interviewer, qualitative researchers can efficiently collect and preliminarily analyze large volumes of data without the need for extensive human effort. We demonstrated that the system performs at a level comparable to experienced human interviewers in real-world setting.

We believe that LM-Interviewer will not only serve as a valuable tool but also expand the boundary of qualitative researches.

## Limitations

We identify two main limitations in LM-Interviewer. (1) Delays during conversation: the reliance on external services, especially large language model APIs, causes delays in question generation, which can reduce the continuity of interviews, leading to decreased effectiveness in information collection. This issue can be mitigated by deploying open-source language models, such as LLAMA 2 (Touvron et al., 2023). (2) Limited interactions. Although interviews are typically conducted through conversations, checklists or forms are still used in specific contexts to improve the efficiency of collecting basic information. We plan to integrate these interaction methods into our system in the future.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62277033). It is also supported by the project from Tsinghua-SPD Bank Joint-Lab. We also acknowledge the support from National Engineering Laboratory for Cyberlearning and Intelligent Technology, Beijing Key Lab of Networked Multimedia, and the Institute for Guo Qiang, Tsinghua University (No.20192920479).

## References

Omolola A Adeoye-Olatunde and Nicole L Olenik. 2021. Research and scholarly methods: Semi-structured interviews. *Journal of the american college of clinical pharmacy*, 4(10):1358–1367.

Shirley Agostinho. 2005. Naturalistic inquiry in e-learning research. *International Journal of Qualitative Methods*, 4(1):62–66.

Loraine Blaxter, Christina Hughes, and Malcolm Tight. 2010. *How to research*. McGraw-Hill Education (UK).

Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Elwood F Holton and Michael F Burnett. 2005. The basics of quantitative research. *Research in organizations: Foundations and methods of inquiry*, pages 29–44.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Steinar Kvale. 2012. *Doing interviews*. Sage.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334.

Annette Lareau. 2021. *Listening to people: A practical guide to interviewing, participant observation, data analysis, and writing it all up*. University of Chicago Press.

Anne M Lillis. 1999. A framework for the analysis of interview data from multiple field research sites. *Accounting & Finance*, 39(1):79–105.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Nuzhat Naz, Fozia Gulab, and Mahnaz Aslam. 2022. Development of qualitative semi-structured interview guide for case study research. *Competitive Social Science Research Journal*, 3(2):42–52.

Fatemeh Rabiee. 2004. Focus-group interview and data analysis. *Proceedings of the nutrition society*, 63(4):655–660.

Kathryn Roulston. 2011. Interview 'problems' as topics for analysis. *Applied linguistics*, 32(1):77–94.

Charles Spearman. 1961. The proof and measurement of association between two things.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sarah J Tracy. 2010. Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative inquiry*, 16(10):837–851.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Annette M Willgens, Robin Cooper, Doles Jadotte, Bruce Lilyea, Cynthia L Langtiw, and Alice Obenchain-Leeson. 2016. How to enhance qualitative research appraisal: Development of the methodological congruence instrument. *The Qualitative Report*, 21(12):2380–2395.

Janet BW Williams. 1988. A structured interview guide for the hamilton depression rating scale. *Archives of general psychiatry*, 45(8):742–747.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

# A Principles in Dialogue Module

Table 2: Principles in Dialogue Module

| Actions | Principles |
|---|---|
| Querying | Only ask one question at a time! This helps keep the interview clear and allows the interviewee to stay focused, making it very important. |
| | Start with general questions, and once you have basic information and a direction for the topic, shift to asking about specific actions, events, or experiences. Focus on specific moments and events rather than general situations. |
| | Keep your questions neutral and open-ended, minimizing yes-or-no type questions. Leave definite or negative questions for the end; do not suggest possible answers to the interviewee. |
| Advancing | Check if the topic has deviated and promptly steer the conversation back to the main subject if necessary. |
| | Remember the interview guidelines and essential questions that need to be asked. Check the progress during the interview and have a basic control over time allocation. |
| | Once a topic has been thoroughly explored, you can return to another topic of interest or move on to the next question. At the end of the interview, ask if everything has been covered sufficiently and bring up any aspects you are particularly interested in. |
| Probing | Actively explore the interviewee's personal feelings, asking questions like "Why do you think that?", "Why do you have these concerns?", "How do you view...?", "How did this make you feel?". |
| | Use probing questions that encourage the interviewee to provide more details about their experiences, such as who, what, when, where, what was said, and how it happened. |
| | When probing, if there are multiple appropriate points of information to inquire about, start from a positive perspective before moving to a negative one. |
| Responding | For interviewees who are reticent, show empathy and understanding, gently coax them to respond; or compliment the interviewee; or switch to discussing other lighter topics to help the interviewee relax; or politely probe further. |
| | Listen attentively, providing responses that could be brief affirmations or repeating parts of what the interviewee has said. |