Clarify User Expertise: Towards Proactive Conversational Agents Tailoring Responses to User Proficiency

Zhihong Cao[1] **Chen Huang**[2,3*]
[1] School of Computing and Data Science, The University of Hong Kong
[2] Institute of Data Science, National University of Singapore
[3] College of Computer Science, Sichuan University
czh040302@163.com, huangc.scu@gmail.com

Abstract

In the context of information seeking, conversational agents are undergoing an evolution from reactive tools to proactive, personalized assistants. A critical aspect of this evolution is the ability to tailor strategic interactions to a user's unique needs and expectations. Unlike existing studies that focus on proactively clarifying query ambiguities, we center on clarifying the user's expertise in order to tailor responses for better user comprehension. We find that existing agents struggle to determine user expertise from queries alone, a limitation that prevents them from dynamically adapting their responses. To address this gap, we introduce PASSING to empower the agent to proactively clarify a user's expertise through targeted inquiries. This is achieved by our *What-to-Ask* and *How-to-Ask* strategies, induced by LLM self-play. Our extensive experiments also show our superiority. We believe that PASSING represents a crucial step towards creating more human-centric conversational agents.

1 Introduction

Conversational agents have become a crucial tool for satisfying user information needs in multi-turn dialogues (Casheekar et al., 2024; Zamani et al., 2023). With the advent of large language models (LLMs), a new paradigm of *Proactive Conversational Agents* has emerged that goes beyond simply reacting to user queries (Deng et al., 2025; Liao et al., 2023; Huang et al., 2026). These agents take the initiative to handle under-specified (Braslavski et al., 2017; Xu et al., 2019) or over-specified user requests (Wu et al., 2023; Min et al., 2019) during the information seeking process, rather than passively following the user's lead. Common proactive behaviors include asking questions to clarify query ambiguity (Zhang et al., 2024b; Chen et al., 2024b) or elicit user preference (Zhang et al., 2018, 2022; Chen et al., 2023). For instance, in response to a broad query like "quantum computing", these agents might ask clarifying question: "Are you asking about the history of quantum computing?". As such, the agent seeks to proactively comprehend user query to deliver a more accurate response.

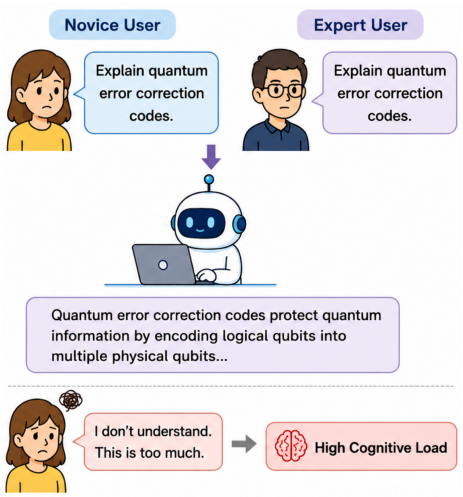

Figure 1: User expertise can vary depending on the specific query. LLM agent should tailor its response based on user expertise.

While current research prioritizes response accuracy, the crucial task of tailoring responses for user comprehension has been largely overlooked. Modern users expect responses that are not only factually correct but also personalized to their individual domain expertise, adjusting the level of detail and complexity accordingly (Dreyfus and Dreyfus, 1980; Schank, 1999; Chen et al., 2024a; Yao et al., 2024; Jeck et al., 2025; Tawfik et al., 2021; Yaghoubzadeh and Kopp, 2017). As illustrated in Figure 1, user expertise can vary significantly depending on the specific query. LLM agents must estimate the user's query-specific expertise level prior to formulating a response. Otherwise, novices, for example, can be overwhelmed by technical details, making them susceptible to

*Corresponding author.

cognitive overload (Dreyfus and Dreyfus, 1980).

To this end, our work centers on clarifying user expertise as a foundational step for generating tailored responses[1]. We begin by experimentally investigating the limitations of current agents in query-specific expertise estimation (Section 3). Our findings reveal a significant expertise inconsistency in LLM-based agents: when presented with the same query, their predictions of user expertise are highly stochastic, with high inconsistency rate. This aligns with previous user study (Palta et al., 2025) and implies that a user's expertise on the specified query is difficult to infer from the query text (Liu et al., 2024). For instance, a computer science PhD student may be a novice in physics but inquiry technical terms like 'quantum computing' without a full grasp of the concept. Therefore, we argue that **proactive conversational agents must probe the user's expertise on a given query via strategic interaction**, a need previously identified (Chen et al., 2024a) but not yet fulfilled.

In this paper, we propose **PASSING**, the first method for ProActive uSer expertiSe probING through targeted inquiries. Recognizing that expertise is often manifested in a user's knowledge scope and reasoning processes (Feltovich and Hoffman, 1997; Palta et al., 2025), PASSING employs a strategy-guided probing mechanism, which utilizes LLM-induced strategies (i.e., *What-to-Ask* and *How-to-Ask*) to direct PASSING in formulating informative inquiries. As illustrated in Figure 2, PASSING leverages these strategies to iteratively pose targeted questions at each turn, obtaining user information on the knowledge scope and reasoning logic based on user's responses. To balance the trade-off between maximizing information gain and minimizing conversational overhead, PASSING stops probing after a pre-defined number of turns. Subsequently, it estimates the user's query-specific expertise and generates an appropriately tailored response. Before putting PASSING into use, PASSING implements an offline self-play simulation to derive these strategies: dialogue histories between PASSING and simulated users are analyzed to extract effective probing strategies, which are then iteratively optimized throughout the self-play process. Notably, distinct from research on ambiguity resolution or preference elicitation, PASSING prioritizes the assessment of user expertise to facilitate better answer comprehension, addressing a vital gap in human-centric AI (Deng et al., 2024).

To evaluate the effectiveness of PASSING, we conduct experiments with various baselines and LLM backbones using various datasets. Experimental results demonstrate that with an average of only 1.2 inquiry turns, PASSING achieves nearly a threefold improvement in the accuracy of user expertise estimation. Finally, we sum up our main contributions as follows.

- We highlight the importance of user expertise estimation for proactive conversational agents, a step toward more human-centric interactions.
- We propose PASSING, featuring accurate expertise estimation through targeted inquiries while minimizing conversational overhead.
- We experimentally demonstrate that existing agents fail to reliably infer user expertise from queries, and validate our effectiveness.

2 Related Work

User Expertise Estimation. A body of theoretical (Dreyfus and Dreyfus, 1980; Schank, 1999) and user-centric (Chen et al., 2024a; Schubert et al., 2013; Palta et al., 2025) research highlights the significant differences between novices and experts in their perceptions and needs. In this case, to generate such personalized response, existing research largely assumes that user expertise is a known information (e.g., via pre-defined user profiles) (Jeck et al., 2025; Sun et al., 2025; Cheng et al., 2024). However, such information is insufficient for determining query-specific expertise, because a user's expertise can vary significantly from one query to the next. While recent methods like ExpertPrompting (Xu et al., 2025) attempt to infer expertise via in-context learning, user studies indicate that LLM-based agents frequently miscalibrate—either underestimating or overestimating the user's knowledge—resulting in reduced satisfaction (Palta et al., 2025). Therefore, proactive agents must be able to probe for this information before generating a response. To fill this gap, we propose PASSING, the first method designed to leverage an agent's proactivity to probe and clarify a user's expertise on the user query through targeted inquiries.

Proactive Conversational Agents. Analogous to general proactive agents that explore an environment to acquire information that improves fu-

[1]Our research scope is confined to the task of query-specific user expertise estimation, a prerequisite for generating tailored responses. Response features required by experts and novices is provided in Appendix A.

ture decision-making (Lu et al., 2025; Guan et al., 2026), proactive conversational agents actively engage users to uncover task-relevant information and facilitate goal completion. These agents are distinguished by their ability to take initiative, steering dialogues toward productive outcomes rather than passively following a user's lead (Deng et al., 2024, 2025). To date, proactive efforts in information seeking have primarily focused on improving response accuracy. This is achieved through strategies such as clarifying query ambiguity (Zhang et al., 2024b; Chen et al., 2024b), eliciting user preferences (Zhang et al., 2018, 2022; Chen et al., 2023), or managing over-specified requests (Wu et al., 2023; Min et al., 2019). However, this intense focus on accuracy has left the crucial task of tailoring responses for user comprehension largely overlooked. User studies indicate a clear demand for agents that first assess a user's knowledge level before providing an answer (Chen et al., 2024a). In response to this need, our work introduces a proactive conversational agent designed specifically to probe for a user's expertise on a given query before generating a response.

2.1 Prediction Stability of Existing Methods

Method	MMLU-Pro		ARC-MCAS	
	Kappa	Unk	Kappa	Unk
Deterministic	100.0	0.00	100.0	0.00
Fully Random	0.00	16.67	0.00	16.67
Direct Prompt	41.25	28.00	30.64	82.80
CoT	28.47	26.40	70.34	89.60
Self-consistency	85.12	32.00	75.35	86.80
Diverse-Aspect	39.91	14.80	47.23	25.60
IDL*	40.31	23.79	15.00	4.40
ExpertPrompting	36.25	25.60	72.00	94.00

Table 1: Prediction stability and unknown rate of existing methods for query-specific user expertise estimation.

3 Preliminary Experiment

This section investigates LLM-based agents can reliably determine user expertise on a given query.

3.1 Experiment Setup

Experiment Overview. A single query can be posed by users with vastly different expertise levels. For instance, a query about *quantum mechanics* could originate from a physics novice, a physics expert, or even a computer science expert who is a novice in quantum mechanics. Considering the scarcity of open-source data labeled with query-specific user expertise, we propose an evaluation methodology focused on prediction stability of an agent's expertise judgments when it is presented with the same query multiple times.

Baselines. As there are no established methods specifically for probing query-specific user expertise, our baselines include two distinct categories. 1) **General LLM-based Agents**: Direct Prompt, Chain-of-Thought (CoT) (Wei et al., 2023), Self-consistency (Wang et al., 2022), and our designed Diverse-Aspect that first prompts the LLM to generate judgments from multiple specific aspects (e.g., terminology use) and then ensembles these aspect-specific judgments to produce the final prediction. 2) **Expertise Estimation**: IDL (Cheng et al., 2024) and ExpertPrompting (Xu et al., 2025). Notably, IDL predicts general user expertise based solely on pre-collected user conversation history[2] and are inherently unable to provide an estimation specific to the given input query. To the best of our knowledge, ExpertPrompting is the only query-specific expertise estimation method. Finally, GPT-4o serves as the LLM backbone.

Dataset. To ground our evaluation in realistic and challenging scenarios, we use questions from two multi-disciplinary datasets as input user queries: MMLU-Pro (Wang et al., 2024) and ARC-MCAS[3] (Clark et al., 2018). For the output classes, we adopt the *Dreyfus Model* (Dreyfus and Dreyfus, 1980), which categorizes user expertise into five levels: Novice, Advanced Beginner, Competent, Proficient, and Expert (detailed in Appendix A). We further augment these five expertise levels with an additional '*Unknown*' category. This results in a six-way classification task where each baseline must classify the user's expertise on the input query into one of these categories. A more comprehensive evaluation is presented in Section 5.

Evaluation Metrics. Given the multi-class classification nature of our task, we measure prediction stability using Fleiss' Kappa (Fleiss, 1971). This metric measures the inter-run agreement for individual predictions. We treat the set of predictions generated under each random seed as an independent "rater" and calculate the Kappa score across all user queries to assess consistency between runs. Additionally, we calculate the proportion of 'Unknown' predictions generated by the model, which serves

[2]History is built from validation set of corresponding data.
[3]Subset of the ARC-Challenge

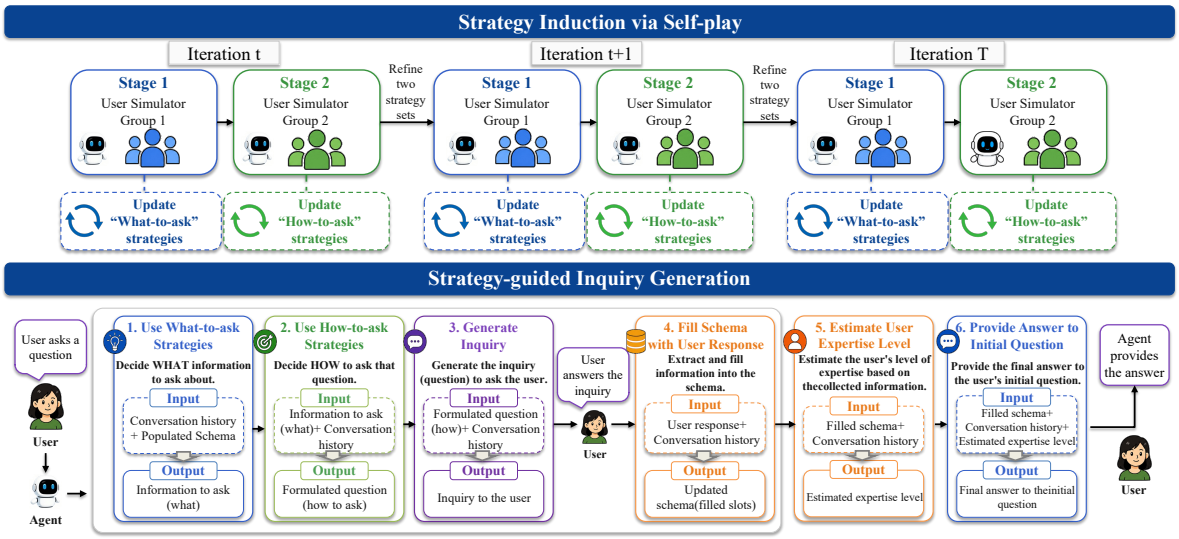

Figure 2: Overview of PASSING. It utilizes LLM-induced strategies (i.e., What-to-Ask and How-to-Ask), obtained through self-play simulation.

as a direct statistical measure of the model's self-assessed inability to determine user query-specific expertise. Refer to Appendix B.5 for details.

Implementation Details. To evaluate the stability of each baseline's judgment, we conducted 5 independent runs for every query. Each of these 5 runs was initialized with the same pre-defined random seeds. More details are in Appendix B.

3.2 Results

LLM agents are unreliable for robustly estimating user expertise specific to a given query. As illustrated in Table 1, while existing methods demonstrate improved stability compared to *fully random* approaches when assessing user query-specific expertise, they still lack consistency compared to the *deterministic* method, often fluctuating in their judgments. The consistently low Kappa scores across methods reveal that even when models do produce a prediction, their judgments are highly inconsistent across runs. Notably, while *Self-consistency* achieves a relatively high Kappa (85.12 on MMLU-Pro, 75.35 on ARC-MCAS), this stability is an artifact of its majority-voting mechanism, which is the aggregation of multiple samples naturally converges to a dominant prediction, rather than reflecting genuine expertise inference. Similarly, the high Kappa of certain methods on ARC-MCAS (e.g., *CoT*: 70.34, *ExpertPrompting*: 72.00) is largely driven by an extremely high Unknown rate (89.60% and 94.00% respectively), indicating that the model repeatedly abstains rather than making meaningful predictions. Importantly, *ExpertPrompting*, which is designed to estimate general user expertise, typically fails to accurately predict query-level expertise. This highlights that any user can copy-paste a jargon-heavy query from the web into an LLM, creating a façade of expertise despite having no actual proficiency in the domain. This motivates our proactive agent to probe the user's query-level expertise.

4 PASSING: Clarify User Expertise

Overview. As illustrated in Figure 2, PASSING estimates a user's query-specific expertise through iterative probing and slot-based state tracking. Given a user query q, PASSING first constructs a query context C_0 using the query and model-generated answer. It then engages the user in a multi-turn probing process guided by predefined probing strategies S. Formally, let $H_t = (p_i, u_i)_{i=1}^{t-1}$ denote the conversation history before turn t, where p_i and u_i represent the system inquiry and user response, respectively, with $u_1 = q$. Let L_t denote the current slot-based expertise schema. At each turn, PASSING generates an inquiry p_t based on the current state (C_0, L_t), receives the user's response u_t, and updates both the conversation state and schema to incorporate newly acquired evidence about the user's expertise. This process repeats until a termination condition is met. Then PASSING produces a query-specific expertise assessment and generates a personalized response accordingly.

Multi-turn Probing and Slot Filling. To capture query-specific expertise, we design an expertise

schema based on the Dreyfus Model (Dreyfus and Dreyfus, 1980; Mangiante and Peno, 2021), where each slot corresponds to a key assessment aspect (Table 2). Following prior work on dialogue state tracking (Cao et al., 2025; Das et al., 2024), PASSING adopts a slot-filling paradigm that incrementally populates the schema using evidence collected during probing. After each user response, relevant slots are updated to reflect newly inferred expertise signals, and the updated schema is subsequently used to guide the next inquiry. To further improve inquiry generation, we incorporate LLM-induced probing strategies, described in Section 4.2.

Query-specific Expertise Assessment. The probing process terminates when either the maximum probing budget T is reached or all schema slots have been populated. PASSING then estimates the user's query-specific expertise using the populated schema together with the accumulated conversation history, and generates a response tailored to both the user's query and estimated expertise level.

4.1 Strategy Induction via Self-play

The key challenge in expertise estimation lies in generating probing questions that effectively reveal user expertise while minimizing cognitive burden. This requires determining both *what* to ask and *how* to ask. To this end, PASSING learns these strategies through offline self-play with user simulators[4] and iteratively refines its probing strategies via trial-and-error. Specifically, strategy induction is performed in two stages using distinct simulators and optimization objectives. The first environment focuses on identifying missing or uncertain knowledge areas, inducing *What-to-Ask* strategies that guide the selection of probing targets. The second environment focuses on maximizing information gain while maintaining user engagement, inducing *How-to-Ask* strategies that govern the phrasing and presentation of inquiries.

Diverse User Simulation. Each simulator is instantiated with a unique persona, including *Big-Five Personality Traits*[5] (Goldberg, 1992) and *Decision-Making Styles*[6] (Scott and Bruce, 1995), to encourage behavioral diversity during interaction. To model varying expertise levels, each simulator is additionally assigned a query and a query-specific expertise label (e.g., *novice*). We further equip the simulator with a query-specific knowledge dependency graph automatically generated by GPT-5, which contains the knowledge required to understand and answer the query. To control the simulator's expertise, we randomly mask different portions of the knowledge dependency graph according to its assigned expertise level. Consequently, novice simulators possess only partial knowledge, whereas expert simulators retain more complete knowledge. This design provides a controllable environment for learning *What-to-Ask* strategies. For *How-to-Ask* strategy induction, the simulator must also exhibit non-cooperative behaviors, as users may refuse to answer poorly phrased or overly demanding questions. Thus, we follow Zhang et al. (2024a) and augment the simulators with refusal strategies and corresponding few-shot examples, enabling them to reject inquiries under appropriate circumstances. This forces PASSING to learn questioning strategies that maximize information gain while maintaining user cooperation.

Algorithm 1 Strategy Induction via Self-play

Require: N: Max simulation sessions, T: Max conversation turns

▷ W: *What-to-Ask*, H: *How-to-Ask*
for $M \in \{W, H\}$ **do**
  Initialize strategy S_0^M
  Initialize user simulators $\mathcal{U}^M$
  for $k = 0$ to $N - 1$ **do**
    Experience $\mathcal{E} \leftarrow \emptyset$
    for all $u \in \mathcal{U}^M$ **do**
      $c \leftarrow \text{SELFPLAY}(S_k^M, u, T)$
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{c\}$
    end for
    $S_{k+1}^M \leftarrow \text{EXTRACT\&REFINE}(S_k^M, \mathcal{E})$
  end for
end for

Self-Play Pipeline. To iteratively optimize probing strategies, we conduct N sessions of offline self-play simulations. We initialize the baseline strategy set S_0 using *Gemini 2.5 Pro*. In each k-th session, PASSING employs the current strategies (S_k) to interact with a user simulator (GPT-4o Mini), aiming to estimate its expertise as described previously. A session terminates upon successful estimation or when the maximum turn limit T is reached. Following each session, we analyze the dialogue logs as experience to induce new strategies, which are then refined to update the set for the next iteration. Refer to Algorithm 1 for details.

- **Strategy Extraction**. Inspired by recent studies in inductive learning, which empowers LLMs to synthesize broader patterns from granular ex-

[4]Refer to Appendix B.4.1 for implementation details

[5]Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism

[6]Directive, Conceptual, Analytical, and Behavioral

Aspects	Descriptions
Knowledge Scope (What they know)	The extent, type, and organization of the knowledge of the person: knowing "facts" (explicit knowledge), knowing when to apply it (applied knowledge), or knowing "how" intuitively (tacit knowledge)
Components (What they perceive)	What types of knowledge or information the person relies on: context-free (rules, abstract principles), rule-based (context-aware rule selection), or situational (contextual, experiential patterns)
Perspective (How they organize info)	How the person filters and prioritizes information: overwhelmed easily, making a deliberate plan to filter noise, or instinctively seeing important information without trying
Action (How they decide)	How decisions are made: analytic (step-by-step reasoning) or intuitive (immediate, holistic response)
Commitment (Emotional connection)	The level of personal agency and emotional investment in the outcome: detached (disengaged, rule-following) or involved (personally invested, responsible for outcome)

Table 2: Structured schema used for estimating user expertise based on the Dreyfus Model.

amples (Cai et al., 2025; de Souza et al., 2025; Huang et al., 2024; Wang et al., 2025; Kim et al., 2025), the accumulated self-play experience is distilled into probing strategies that generalize across users and queries. For each self-play stage, we collect simulation dialogue logs together with the probing strategies employed during interaction. We then analyze these trajectories from two complementary perspectives: (1) the accuracy of the final expertise assessment, and (2) the information gain obtained at each probing turn. Using the resulting dialogue trajectory and assessment outcome as context, an LLM is prompted to identify why a particular strategy contributed to successful or failed expertise estimation, or why a probing turn yielded high or low information gain. The extracted insights are subsequently distilled into strategies in a structured form: *Probe [TARGET] to elicit information regarding [ASPECT], because [RATIONALE].*

- **Strategy Refinement**. Since a single session comprises multiple turns, numerous raw strategies are extracted. To prevent these strategies from becoming overly specific to the query or its associated knowledge points, and to avoid redundancy, we employ an LLM-based semantic clustering approach, together with the strategies from previous session. Strategies within the same cluster are then abstracted to derive high-level probing tactics for use in subsequent simulations.

Note that we use the same structured template for both strategy types to control for sentence-structure confounds while allowing their content to differ. Each template specifies the probing target, the expertise-related information sought, and the probe's utility. What-to-Ask strategies select diagnostic targets, whereas How-to-Ask strategies determine how those targets are elicited conversationally.

4.2 Strategy-guided Inquiry Generation

Given the populated schema L_t and conversation history H_t, PASSING generates the next inquiry in a two-stage manner. First, a *What-to-Ask* strategy is selected based on the current schema and conversation history to determine which aspect of the user's expertise should be further probed. Conditioned on the selected strategy and conversation history, a *How-to-Ask* strategy is then chosen to determine how the inquiry should be phrased. The final inquiry is then generated based on the formulated question and conversation history. After receiving the user's response, PASSING extracts and fills information into the schema based on the user response and conversation history, resulting in an updated schema with filled slots. This process continues until either all schema slots have been populated or the max conversation turn is reached. At that point, PASSING uses the filled schema and conversation history to estimate the user's query-specific expertise level. The estimated level, together with the filled schema and conversation history, is then used to generate a personalized response. Notably, the entire pipeline is implemented through the zero-shot capabilities of LLMs, including strategy selection and expertise estimation. While these components could alternatively be replaced by dedicated trainable modules, the lack of high-quality training data and our exploratory nature of query-specific expertise estimation motivate us to adopt a training-free design in this work. We leave more complex trainable formulations to future research.

LLM Backbone	Method	MMLU-Pro						ARC-MCAS					
		Acc.↑	Kappa↑	FA↑	Comp.↑	US↑	Unk.↓	Acc.↑	Kappa↑	FA↑	Comp.↑	US↑	Unk.↓
Evaluation with LLM Simulators													
GPT-4o	IDL (*the best baseline*)	20.0	40.31	3.75	3.62	-	23.79	20.0	15.00	**4.59**	3.42	-	4.40
	Direct Prompt	16.0	41.25	3.84	3.69	-	28.00	20.0	30.64	**4.59**	3.67	-	82.80
	PASSING	**77.1**	**64.90**	3.85	3.83	**4.01**	**0.0**	**87.4**	71.28	**4.59**	**3.87**	**4.18**	**0.0**
	PASSING w/o What-to-ask	68.6	64.43	3.81	**3.87**	3.96	**0.0**	76.8	67.26	4.45	3.86	4.08	**0.0**
	PASSING w/o How-to-ask	**77.1**	64.24	**3.93**	3.86	3.97	**0.0**	73.7	**71.31**	4.58	3.84	4.10	**0.0**
DeepSeek V4-Pro	IDL (*the best baseline*)	16.0	45.97	**4.33**	3.08	-	2.0	20.0	55.70	4.49	3.62	-	**0.0**
	Direct Prompt	20.0	**58.10**	4.26	3.69	-	**0.0**	22.0	56.77	**4.50**	3.77	-	**0.0**
	PASSING	**77.1**	51.28	4.08	**3.88**	**4.45**	**0.0**	**78.9**	52.05	**4.50**	**3.94**	**4.54**	**0.0**
	PASSING w/o What-to-ask	**77.1**	50.84	3.99	3.87	4.24	**0.0**	68.4	56.54	4.46	3.92	4.44	**0.0**
	PASSING w/o How-to-ask	65.7	54.80	3.91	3.81	4.17	**0.0**	72.6	**57.54**	**4.50**	3.90	4.33	**0.0**
Evaluation with Human Participants													
GPT-4o	IDL	20	40.31	3.91	3.41	3.37	23.79	20	15.00	3.73	3.47	3.55	4.40
	PASSING	**68.7**	**52.95**	**4.52**	**4.62**	**4.56**	**0.0**	**76.7**	**76.65**	**4.35**	**4.74**	**4.58**	**0.0**
DeepSeek V4-Pro	IDL	16	45.97	3.36	3.26	3.31	2.0	20	**55.70**	3.95	3.35	3.31	**0.0**
	PASSING	**67.3**	**56.98**	**4.38**	**4.56**	**4.43**	**0.0**	**56.0**	38.70	**4.64**	**4.55**	**4.58**	**0.0**

Table 3: Main results. Best results are in bold, and second-best results are underlined.

5 Experiments

5.1 Experimental Setup

We adopt the setup from Section 3, with the following key additions to expand our evaluation:

Baselines & Backbones. We further involve two LLM backbones (GPT-4o and DeepSeek-V4-Pro) for more comprehensive evaluation. We selected IDL and Direct Prompt for the main comparison. IDL is the strongest baseline overall, with a low rejection rate and stable Kappa scores, while Direct Prompt serves as the most fundamental baseline, representing an agent without explicit user-expertise modeling.

Multi-turn Conversations. To facilitate the multi-turn conversations required for expertise probing, we employ a two evaluation strategy involving both user simulators and real human participants (cf. Appendix B.4). For simulator-based evaluation, we adopt the same simulation framework used in *How-to-Ask* strategy induction. To avoid evaluation bias, however, the evaluation simulators are entirely disjoint from those used during strategy induction, with no overlap in personas or refusal strategies. For human evaluation, we recruit participants to interact directly with PASSING under more realistic settings. Additionally, since expertise levels are defined for each user–query pair rather than for the query alone[7], for each query, participants are assigned a target expertise level and instructed to role-play accordingly. To ensure consistency with the assigned query-specific expertise, we provide a knowledge base specifying the concepts expected to be known at that expertise level.

Evaluation Metrics. Beyond prediction stability, we evaluate Prediction Accuracy (Acc.) for query-specific expertise estimation using the ground-truth expertise labels available in our multi-turn conversations. We additionally assess generated responses in terms of factual accuracy (FA), comprehensibility (Comp.), which evaluates whether responses are appropriately adapted to the user's expertise level; and user satisfaction (US), which measures the quality of the overall interaction and the extent to which our probing strategy minimizes user burden. Furthermore, we conduct both human and LLM-based evaluations to assess FA, Comp. and US. Refer to Appendix B.6 for details.

5.2 Main Results

PASSING consistently establishes superior performance across various LLM backbones and datasets. As shown in Table 3, PASSING consistently outperforms the strongest baseline across different LLM backbones and evaluation benchmarks, achieving an average improvement of +288% in query-level expertise estimation accuracy. Benefiting from more accurate estimation of users' expertise levels, PASSING is able to generate responses that are better tailored to users' knowledge boundaries, substantially improving answer comprehensibility (+4.1%). Importantly, despite introducing additional probing interactions with users, PASSING maintains factual accuracy comparable to existing baselines, demonstrating that proactive probing

[7]For example, for the same query on quantum mechanics, a computer science PhD student may be a novice, whereas a physics graduate student may be proficient.

does not compromise response correctness. These results suggest that proactively eliciting user expertise provides an effective mechanism for delivering personalized responses.

PASSING demonstrates relatively reliable performance despite the inherent stochasticity of LLM-based probing. Compared with strong baselines such as IDL and Direct Prompt, PASSING reduces the unknown response rate to nearly zero while maintaining relatively high prediction stability (+42.5%, on average), indicating that the observed performance gains are not driven by random fluctuations. We also observe that PASSING does not always achieve the highest stability scores. We attribute this to the inherent stochasticity of LLMs in both selecting probing strategies across both the what-to-ask and how-to-ask stages, and estimating user expertise under the predefined schema. Although such variance does not diminish the overall advantage of PASSING over existing methods, it highlights a promising direction for future work: improving system stability through dedicated modules for each stage, especially when supervised training data are available.

Probing strategies in PASSING improve user modeling and help maintain user satisfaction. Beyond expertise estimation accuracy, PASSING achieves promising user satisfaction scores across different datasets and LLM backbones. Ablation results further show that removing either the *What-to-Ask* or *How-to-Ask* strategy leads to performance degradation, highlighting the importance of both selecting appropriate probing content and generating effective probing interactions. In particular, removing the *How-to-Ask* strategy consistently decreases user satisfaction, suggesting that the manner of questioning substantially affects user experience. Interestingly, removing the *What-to-Ask* strategy also harms satisfaction. We find that, without explicit guidance on probing content, the generated questions become less controllable and occasionally mismatch user expertise levels: overly difficult questions tend to frustrate users, while simpler questions are generally better tolerated.

Human evaluation further validates the effectiveness of PASSING. As shown in Table 3, PASSING consistently outperforms IDL across both two backbones under human participants and human evaluation. In particular, PASSING achieves substantially higher expertise estimation accuracy and user satisfaction while reducing the unknown response rate to nearly zero. Meanwhile, the factual accuracy and completeness of generated responses remain competitive, demonstrating that proactive probing can effectively improve personalized interactions without sacrificing response quality. Following Bhagwatkar et al. (2025), Zhang et al. (2025), and Irawan et al. (2025), we further examine the consistency between human evaluators and LLM-based evaluators on human interaction data using Gwet's AC2 score (Gwet, 2008). The human–human AC2 reaches an average of 0.72, confirming consistent annotation quality across participants. The human–LLM AC2 reaches an average of 0.79, indicating strong agreement between human judgments and LLM-based evaluation result. Detailed reliability computation and results are provided in Appendix B.8.

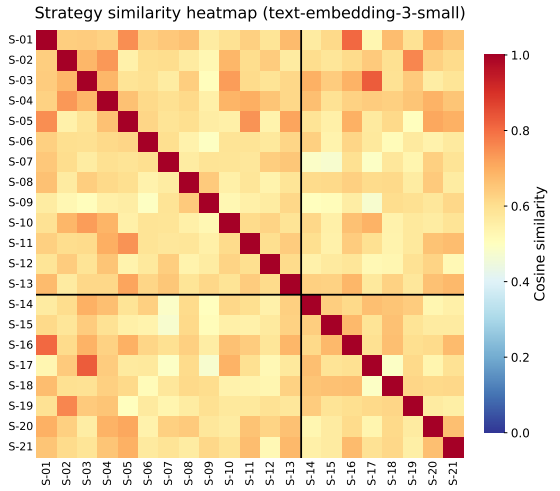

Figure 3: Pairwise similarity among induced strategies (13 What-to-ask and 8 How-to-ask strategies in total)

5.3 In-depth Analysis

Induced strategies reveal a functional rather than topical organization. We further examine whether the induced What-to-Ask and How-to-Ask strategies occupy distinct regions in embedding space. Using the OpenAI *text-embedding-3-small*, we encode all induced strategies and compute pairwise cosine similarities[8] (Figure 3). The two categories are not semantically separated: the average intra-class similarities are 0.608 for *What-to-Ask* and 0.604 for *How-to-Ask*, while the inter-class similarity is 0.604, with a silhouette score of 0.006. This suggests that the two categories differ primarily in functional role rather than topical content: *What-to-Ask* strategies specify diagnostic targets, whereas *How-to-Ask* strategies operationalize these

[8]More details are in Appendix B.7.

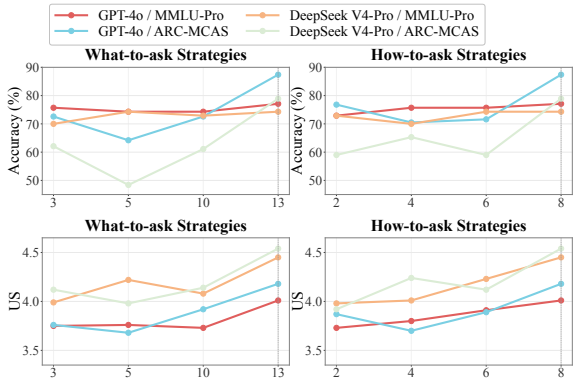

Figure 4: Effect of the number of probing strategies on expertise estimation accuracy and user satisfaction.

targets through conversational tactics.

Maintaining sufficient strategies ensures the effectiveness of proactive probing. We further investigate the impact of the number of probing strategies on both expertise estimation accuracy and user satisfaction in PASSING. Specifically, given 13 *What-to-Ask* strategies and 8 *How-to-Ask* strategies, we randomly sample k strategies from one category while keeping the other category complete, and evaluate the resulting performance. As shown in Figure 4, increasing the number of strategies generally improves performance, suggesting that richer probing diversity helps the model better capture user expertise and interaction preferences.

6 Conclusion

We introduce PASSING, a proactive probing method that enables LLM agents to infer query-specific user expertise before generating responses. This is achieved through our LLM-induced two strategy sets. Generally speaking, our findings suggest that effective personalization should not solely depend on passively modeling users from limited observations, but also on actively acquiring missing information through interaction. In this sense, proactive probing serves as a lightweight yet scalable mechanism in real-world interactions. We hope this work can inspire future research on interaction-centric personalization, adaptive information acquisition, and more human-aware LLM agents.

Limitations

As with prior studies on proactive agents by prompting LLMs, the performance of PASSING may be influenced by prompt design. Prompt sensitivity remains an inherent challenge for LLM-based systems, and exploring more robust prompting or training-based alternatives constitutes an important direction for future work. Another limitation of this work is that we do not systematically investigate the impact of self-play hyper-parameters, such as the number of self-play sessions or the maximum conversation turns, on the quality of induced probing strategies. Although we analyze the effect of the number of strategies in our experiments, the strategy induction process itself remains underexplored. Understanding how different self-play configurations influence strategy diversity, quality, and downstream personalization performance is an important direction for future work. Finally, although What-to-Ask and How-to-Ask strategies serve distinct roles, they inevitably exhibit semantic overlap. As an early study of query-specific expertise probing, we leave systematic disentanglement to future work. One promising approach is to condition second-stage extraction on first-stage strategies and explicitly discourage semantic redundancy.

Ethical Considerations

Similar to prior work (Zhang et al., 2024b; Chen et al., 2024b; Zhang et al., 2018, 2022; Chen et al., 2023), the primary experiments in this paper rely on LLM-based user simulators for large-scale evaluation. In addition, we conduct supplementary human evaluation experiments with voluntary participants. All participants were informed of the experimental setting and engaged in standard question-answering interactions without sensitive topics or psychologically stressful tasks. Therefore, we believe the study poses minimal ethical risk.

LLM Usage

LLMs are used in this work as the backbone models for conversational agents and evaluators in our experiments. In addition, LLMs are also used for language polishing during paper writing.

References

Rishika Bhagwatkar, Syrielle Montariol, Angelika Romanou, Beatriz Borges, Irina Rish, and Antoine Bosselut. 2025. Cave: Detecting and explaining commonsense anomalies in visual environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27110–27151.

Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean

exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 345–348.

Chengkun Cai, Xu Zhao, Haoliang Liu, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, Jenq-Neng Hwang, and Lei Li. 2025. The role of deductive and inductive reasoning in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16780–16790.

Jieming Cao, Chen Huang, Yanan Zhang, Ruibo Deng, Jincheng Zhang, and Wenqiang Lei. 2025. Breaking the stigma! unobtrusively probe symptoms in depression disorder diagnosis dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 182–200, Albuquerque, New Mexico. Association for Computational Linguistics.

Avyay Casheekar, Archit Lahiri, Kanishk Rath, Kaushik Sanjay Prabhakar, and Kathiravan Srinivasan. 2024. A contemporary review on chatbots, ai-powered virtual conversational agents, chatgpt: Applications, open challenges and future research directions. *Computer Science Review*, 52:100632.

John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024a. Learning agent-based modeling with llm companions: Experiences of novices and experts using chatgpt & netlogo chat. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. 2024b. STYLE: Improving domain transferability of asking clarification questions in large language model powered conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10633–10649, Bangkok, Thailand. Association for Computational Linguistics.

Yue Chen, Dingnan Jin, Chen Huang, Jia Liu, and Wenqiang Lei. 2023. Travel: Tag-aware conversational faq retrieval via reinforcement learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3861–3872.

Chuanqi Cheng, Quan Tu, Wei Wu, Shuo Shang, Cunli Mao, Zhengtao Yu, and Rui Yan. 2024. "in-dialogues we learn": Towards personalized dialogue without pre-defined profiles through in-dialogue learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10422, Miami, Florida, USA. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. 2024. S3-dst: Structured open-domain dialogue segmentation and state tracking in the era of llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14996–15014.

João Pedro Gandarela de Souza, Danilo Carvalho, and André Freitas. 2025. Inductive learning of logical theories with llms: A expressivity-graded analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23752–23759.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.

Stuart E Dreyfus and Hubert L Dreyfus. 1980. A five-stage model of the mental activities involved in directed skill acquisition. Technical report.

Paul J Feltovich and Robert R Hoffman. 1997. *Expertise in context*. AAAI Press Menlo Park, CA.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Zhizhao Guan, Chen Huang, Ziming Liu, Hongru Liang, Wenqiang Lei, See-Kiong Ng, Tat-Seng Chua, and Anthony G Cohn. 2026. Clearing the fog: Towards installing and refining proactive exploration capabilities in llm agents. *Preprint*, arXiv:2608.14339.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Chen Huang, Zitan Jiang, Zou Changyi, Wenqiang Lei, and See-Kiong Ng. 2026. Towards proactive information probing: Customer service chatbots harvesting value from conversation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 28768–28792, San Diego, California, United States. Association for Computational Linguistics.

Chen Huang, Yiping Jin, Ilija Ilievski, Wenqiang Lei, and Jiancheng Lv. 2024. Araida: Analogical reasoning-augmented interactive data annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10660–10675.

Patrick Amadeus Irawan, Genta Indra Winata, Samuel Cahyawijaya, and Ayu Purwarianti. 2025. Towards efficient and robust VQA-NLE data generation with large vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4323–4340, Abu Dhabi, UAE. Association for Computational Linguistics.

Jakub Jeck, Florian Leiser, Anne Hüsges, and Ali Sunyaev. 2025. Tell-me: Toward personalized explanations of large language models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Namyoung Kim, Kai Tzu-iunn Ong, Yeonjun Hwang, Minseok Kang, Iiseo Jihn, Gayoung Kim, Minju Kim, and Jinyoung Yeo. 2025. Principles: Synthetic strategy memory for proactive dialogue agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21329–21368.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 3452–3455.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open- and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 452–461, New York, NY, USA. Association for Computing Machinery.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, and 1 others. 2025. Proactive agent: Shifting llm agents from reactive responses to active assistance. In *International Conference on Learning Representations*, volume 2025, pages 47431–47457.

Elaine M Silva Mangiante and Kathy Peno. 2021. *Teaching and learning for adult skill acquisition: applying the Dreyfus and Dreyfus model in different fields*. IAP.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Shramay Palta, Nirupama Chandrasekaran, Rachel Rudinger, and Scott Counts. 2025. Speaking the right language: The impact of expertise alignment in user-ai interactions. *arXiv preprint arXiv:2502.18685*.

Roger C Schank. 1999. *Dynamic memory revisited*. Cambridge University Press.

Christiane C Schubert, T Kent Denmark, Beth Crandall, Anna Grome, and James Pappas. 2013. Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department. *Annals of emergency medicine*, 61(1):96–109.

Susanne G Scott and Reginald A Bruce. 1995. Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement*, 55(5):818–831.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296.

Andrew A Tawfik, Jessica D Gatewood, Jaclyn J Gish-Lieberman, and Charles W Keene. 2021. Exploring the differences between experts and novices on inquiry-based learning cases. *Journal of Formative Design in Learning*, 5(2):97–105.

Borui Wang, Kathleen McKeown, and Rex Ying. 2025. Dystil: Dynamic strategy induction with large language models for reinforcement learning. *arXiv preprint arXiv:2505.03209*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. InSCIt: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2025. Expertprompting: Instructing large language models to be distinguished experts. *Preprint*, arXiv:2305.14688.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Ramin Yaghoubzadeh and Stefan Kopp. 2017. Enabling robust and fluid spoken dialogue with cognitively impaired users. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Saarbrücken, Germany. Association for Computational Linguistics.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. README: Bridging medical jargon and lay understanding for patient education through data-centric NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629, Miami, Florida, USA. Association for Computational Linguistics.

Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell, and Susan T. Dumais. 2020. Analyzing and learning from user interactions for search clarification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1181–1190, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, and 1 others. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456.

Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2025. LLMTaxo: Leveraging large language models for constructing taxonomy of factual claims from social media. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19627–19641, Vienna, Austria. Association for Computational Linguistics.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024a. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 424–444, Miami, Florida, USA. Association for Computational Linguistics.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024b. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. 2022. Multiple choice questions based multi-interest policy learning for conversational recommendation. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2153–2162, New York, NY, USA. Association for Computing Machinery.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 177–186, New York, NY, USA. Association for Computing Machinery.

A Background on User Expertise

A.1 Dreyfus Model

The Dreyfus Model (Dreyfus and Dreyfus, 1980) describes skill acquisition as a progression through five stages, each characterized by distinct differences across five dimensions: knowledge scope, components, perspective, action, and commitment. Novices possess only explicit factual knowledge, rely entirely on context-free rules, are easily overwhelmed by complex information, reason purely step-by-step, and feel little personal investment in outcomes. Advanced Beginners have slightly broader knowledge but still treat all information as equally relevant and struggle to apply rules to real-world situations. Competent performers mark a turning point: they consciously plan and organize information to filter out noise, and begin to feel personally responsible for their decisions. Proficient users have accumulated rich experiential knowledge that allows them to instantly recognize

what matters in a situation without deliberate effort, though they still verify their responses through analytic reasoning. Experts have fully internalized their domain knowledge, rather than recalling facts or following rules, they perceive situations holistically and respond immediately and intuitively.

A.2 Response Features Required by Experts and Novices

The substantial differences across Dreyfus levels imply that responses must be tailored accordingly. Novices require short, jargon-free responses that explicitly explain basic concepts in simple language, as they can only process context-free facts and are easily overwhelmed by complex information. Experts, by contrast, expect dense, conclusion-first, telegraphic responses that use domain jargon without explanation and omit all foundational scaffolding.

B Experiment Details

We use Python 3.10.12, NumPy 1.26.4, and scikit-learn 1.3.2 for implementation and evaluation, including the computation of Fleiss' Kappa (Fleiss, 1971). For preliminary experiments, we randomly sample 50 user queries from the MMLU-Pro and ARC-MCAS datasets. For the main experiments, we use the full datasets.

B.1 Details of Datasets

MMLU-Pro (Wang et al., 2024) is a challenging multi-task language understanding benchmark spanning 14 academic disciplines. We use 70 questions from the test split for offline self-play strategy induction, and sample 50 questions from the validation split for evaluation to avoid data leakage.

ARC-MCAS (Clark et al., 2018) is a subset of the ARC-Challenge benchmark, consisting of science questions from the Massachusetts Comprehensive Assessment System (MCAS). We sample 95 questions from the test split for evaluation.

B.2 Implementation Details of Baselines

Direct Prompt. The LLM is directly prompted to predict the user's query-specific expertise level given the input query, without any additional reasoning or ensemble steps.

Chain-of-Thought (CoT). The LLM is prompted to reason step-by-step before predicting the user's expertise level (Wei et al., 2023).

Self-consistency. Multiple predictions are sampled independently under different random seeds, and the final prediction is determined by majority voting (Wang et al., 2022).

Diverse-Aspect. The LLM first identifies multiple aspects of the query that are relevant to expertise assessment (e.g., terminology use, reasoning style). For each aspect, an independent expertise judgment is generated. These aspect-level assessments are then integrated into a final cohesive expertise prediction.

IDL. The LLM is provided with pre-collected dialogue history from the same user as reference. It extracts structured knowledge triples from the history to form a persona constraint, and uses this constraint to assess the user's expertise level on the new query (Cheng et al., 2024).

ExpertPrompting. The LLM is prompted with few-shot examples to assess the user's expertise level directly from the input query, and generates a tailored response accordingly (Xu et al., 2025).

B.3 Implementation Details of PASSING

As emphasized in prior conversational AI research, agent-initiated probing must account for both users' willingness to respond and the interaction burden it imposes. Although real users do engage with LLM agents and system-initiated questions in realistic information-seeking scenarios (Deng et al., 2023; Zamani et al., 2020; Chiang et al., 2024), demonstrating the practical relevance of this research setting, such engagement should not be assumed to be unconditional or cost-free. Our problem formulation should therefore explicitly model users' willingness to answer probing questions (i.e., How-to-Ask and What-to-Ask). In particular, our PASSING aims to provide responses tailored to the user's query-specific expertise while maintaining user satisfaction. This is achieved by inducing two complementary strategy sets: What-to-Ask selects the most diagnostic expertise evidence, while How-to-Ask phrases the probe in a natural and low-burden way.

To enhance the induction of *What-to-ask strategies*, we further involve a knowledge dependency graph into the PASSING's prompt. Given a user query q and the model-generated answer[9] a, PASSING employs a structured approach to identify and represent the prerequisite knowledge required for comprehending the answer via prompts. During this process, key domain-specific concepts, includ-

[9]The specific technique used for answer generation (e.g., retrieval augmentation or asking clarifying questions) is not the primary focus of our work.

ing terminology entities and actions, are extracted as nodes, while their semantic relationships form the edges. Finally, a knowledge dependency graph G is constructed. Note that each node is additionally assigned a *min-level* attribute indicating the minimum expertise level required to understand that concept; this attribute is used by the user simulator to determine which concepts are known or unknown to a simulated user, but is masked from PASSING during probing to prevent information leakage. In particular, our masking mechanism performs content-based masking according to concept difficulty. Each node in the knowledge dependency graph has a min-level attribute indicating the minimum Dreyfus level required to understand that concept. For a simulator assigned level l, we retain nodes whose min-level is no higher than l and mask the remaining nodes. Thus, a Novice retains only Novice-level concepts, an Advanced Beginner retains concepts from the first two levels, and so forth, while an Expert retains all concepts. Since graph size and level distribution vary by query, there is no fixed retention ratio for each level. The min-level attribute is visible only to the user simulator and hidden from PASSING during probing to prevent information leakage.

B.4 Implementation Details of User Simulators and Human Participants

B.4.1 User Simulators for Strategy Induction

We detail the expertise Profile for different type of user as follows

- The Novice operates with a Knowledge Scope limited to explicit facts and definitions, lacking any deep understanding of "how" things work in practice. They perceive Components strictly as context-free features, relying entirely on rigid rules and abstract principles regardless of the situation. Lacking a functional Perspective, they are often unable to filter information effectively, forcing them to adhere strictly to instructions to avoid becoming overwhelmed. Their Action is purely analytic, characterized by slow, step-by-step reasoning to execute instructions. Consequently, their Commitment remains detached; they view themselves as simply following the rules and feel little personal agency or responsibility for the outcome.

- The Advanced Beginner has expanded their Knowledge Scope slightly but still relies heavily on explicit facts. While they begin to recognize some situational Components alongside context-free rules, they lack the experience to distinguish importance. As a result, their Perspective is flawed; they treat all information as equally relevant, which leads them to be easily overwhelmed by the noise of a complex situation. Their Action remains analytic and deliberate as they struggle to apply guidelines to real-world scenarios. Like the Novice, their Commitment is largely detached, as they are focused on surviving the task rather than owning the result.

- The Competent performer marks a major shift in Perspective. Faced with a complex situation, they do not get overwhelmed; instead, they consciously make a deliberate plan to organize information and filter out noise. Their Knowledge Scope is now organized enough to handle this structuring. While they perceive both context-free and situational Components, they prioritize them based on a hierarchical goal. Their Action is still analytic—they solve problems by reasoning through their plan step-by-step. However, their Commitment shifts to being involved; because they chose the plan, they feel a sense of personal responsibility and emotional investment in the outcome.

- The Proficient user possesses a vast Knowledge Scope of tacit knowledge derived from experience. Their perception of Components is now dominated by situational patterns rather than rules. Crucially, their Perspective has evolved from "making a plan" to "seeing the picture"; they instinctively distinguish important information from noise without conscious effort. Despite this intuitive perception, their Action remains analytic; they effectively diagnose the problem instantly but still rely on step-by-step reasoning to calculate the best specific response. Their Commitment is deeply involved in the perception of the problem, though they may remain analytically detached when executing the solution.

- The Expert operates with a deep, tacit Knowledge Scope where knowing "that" has completely transformed into knowing "how." They perceive Components entirely through situational patterns, reading the environment holistically. Their Perspective is immediate and instinctive; important information jumps out at them, and they are never overwhelmed. The defining difference lies in

their Action: it shifts from analytic to intuitive. They no longer rely on step-by-step reasoning to decide; the correct response arises immediately and fluidly. Their Commitment is fully involved, acting with a level of agency where the person and the task are effectively merged.

What-to-Ask vs. *How-to-Ask* Simulator Design:

- While both simulators instantiate users according to the five Dreyfus profiles described above, they differ in design. The *What-to-Ask* simulator injects only the expertise-level profile and knowledge boundaries (known/unknown concepts), and the simulator's sole task is to generate a response consistent with the cognitive style of that level; responses are output as plain text with no explicit decision process, yielding five fixed behavioral distributions. The *How-to-Ask* simulator builds on this foundation by overlaying a personality trait drawn from the Big Five model (Goldberg, 1992) and a decision-making style drawn from four archetypes (Scott and Bruce, 1995), allowing users at the same expertise level to exhibit markedly different communicative styles. It further introduces two additional mechanisms: a level-specific refusal pattern that characterizes how users at each expertise level typically decline to answer, and an explicit answer/refuse decision step that requires the simulator to weigh reasons for answering against reasons for refusing before producing a structured output containing both a decision and a response. The resulting training dialogues cover uncooperative behaviors including deflection, minimal engagement, and outright refusal, forcing the how-to-ask strategies to learn how to adapt their phrasing when users are not forthcoming.

B.4.2 User Simulators for Experimental Evaluation

For the main experiments, we employ GPT-4o Mini as the user simulator. Each simulator instance is initialized with a pre-defined expertise profile comprising: (1) an expertise level drawn from the five Dreyfus levels; (2) a persona description including personality trait based on the Big Five model (Goldberg, 1992) and decision-making style (Scott and Bruce, 1995); (3) a set of known and unknown concepts derived from the knowledge dependency graph. To better simulate realistic user behavior, each expertise level is assigned a corresponding refusal pattern that allows the simulator to autonomously decide whether to answer or refuse a given inquiry. The simulator strictly respects its knowledge boundaries by never displaying knowledge of unknown concepts, enabling diverse and realistic simulation of users across different expertise levels.

B.5 Implementation Details of Evaluation Metrics

Prediction Stability (Kappa). We measure prediction stability using Fleiss' Kappa (Fleiss, 1971), defined as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e},$$

where $\bar{P}$ denotes the mean observed agreement across all queries, and $\bar{P}_e$ denotes the expected agreement under chance. To quantify cross-run prediction consistency, we treat predictions produced under different random seeds as independent raters. The resulting Kappa score reflects how stable the method is across repeated runs.

Factual Accuracy (FA). We employ GPT-5 as a third-party evaluator to assess the factual accuracy of the tailored responses. The evaluation covers four dimensions: core answer correctness, detail correctness, no factual error, and no misleading, each scored on a 1–5 scale. The final FA score is the average across all four dimensions.

Comprehension (Comp.). Comprehension is evaluated by the user simulator (GPT-4o Mini) from the perspective of the simulated user, initialized with a pre-defined expertise profile including known and unknown concepts. The simulator assesses four dimensions: clarity, level match, completeness, and accessibility, each scored on a 1–5 scale. The final Comp. score is the average across all four dimensions.

User Satisfaction (US). We employ GPT-5 as a third-party evaluator to assess user satisfaction based on the full conversation history. The evaluation covers five dimensions: comfort, feeling understood, willingness to engage, conversation naturalness, and overall satisfaction, each scored on a 1–5 scale. The final US score is the average across all five dimensions.

B.6 Implementation Details of Human Evaluation

B.6.1 Role-playing Protocol

We recruited ten graduate students specializing in computer science and artificial intelligence to par-

ticipate in the human evaluation. All participants voluntarily took part in the study.

Each participant role-played users at five expertise levels defined by the Dreyfus Model: Novice, Advanced Beginner, Competent, Proficient, and Expert. The expertise profiles used for role-playing follow the same Dreyfus-based descriptions as those used in our user simulators for strategy induction, as detailed in Appendix B.4.1. Each participant completed 10 conversations per level, yielding 50 conversations per participant. Experiments were conducted on two datasets, MMLU-Pro and ARC-MCAS, and two LLM backbones, GPT-4o and DeepSeek-V4-Pro.

To ensure consistent role-playing behavior across participants, we provided participants with the expertise-level guide described in Appendix B.4.1, together with representative example responses. This design allowed participants to focus on controlling the reasoning style, uncertainty, and response granularity associated with each target level.

After each conversation, participants evaluated the system's tailored response on three dimensions using 5-point Likert scales. Factual Accuracy (FA, 4 items) measures factual consistency between the system response and a provided reference answer, independent of expertise-level adaptation. Comprehensibility (Comp, 4 items) assesses whether the response is clear and understandable for the role-played user level. User Satisfaction (US, 5 items) measures whether the response is appropriate, helpful, and well aligned with the participant's assigned expertise level.

B.7 Implementation Details of Experiment

Semantic analysis of induced strategies. We provide additional details for the semantic analysis in Section 5.3. We encode all 21 induced strategies using an OpenAI-compatible embedding endpoint based on text-embedding-3-small, obtaining a 1536-dimensional embedding for each strategy. The strategy set contains 13 *What-to-Ask* strategies and 8 *How-to-Ask* strategies. We compute pairwise cosine similarities and report intra-class similarity, inter-class similarity, the diversity gap, and the silhouette score under the *What-to-Ask*/*How-to-Ask* labeling.

The diversity gap is defined as:

$$\Delta = \frac{\mathrm{Sim}_{\textit{What-to-Ask}} + \mathrm{Sim}_{\textit{How-to-Ask}}}{2} - \mathrm{Sim}_{\mathrm{inter}},$$

Metric	Value
What-to-Ask intra-class similarity	0.608
How-to-Ask intra-class similarity	0.604
Inter-class similarity	0.604
Diversity gap Δ	0.003
Silhouette score	0.006

Table 4: Semantic similarity statistics for induced strategies.

where $\mathrm{Sim}_{\textit{What-to-Ask}}$ and $\mathrm{Sim}_{\textit{How-to-Ask}}$ are the average within-category cosine similarities excluding diagonal self-similarities, and $\mathrm{Sim}_{\mathrm{inter}}$ is the average cross-category cosine similarity.

The closest cross-category pairs are S-03/S-17 (0.825), S-01/S-16 (0.807), and S-02/S-19 (0.759), corresponding to conceptual distinction, rationale probing, and concrete application, respectively. These nearest-neighbor pairs support the interpretation that *How-to-Ask* strategies specify *How-to-Ask* questions that elicit evidence for diagnostic targets defined by *What-to-Ask* strategies.

B.8 Human Evaluation Reliability

Setup. To assess the reliability of the human evaluation described in Section B.6.1, we compute inter-rater agreement from two perspectives. First, we measure *human–human* agreement among the human participants to verify that the evaluation criteria were applied consistently. Second, we measure *human–LLM* agreement between the human ratings and the LLM-based evaluator to examine whether the automatic evaluator aligns with human judgment. Reliability is computed separately for the three post-conversation evaluation dimensions: factual accuracy (FA), comprehensibility (Comp.), and user satisfaction (US).

Metric and computation. We use ordinal-weighted **Gwet's AC2** (Gwet, 2008) as the reliability metric. AC2 is a chance-corrected agreement coefficient that compares the observed agreement among raters with the agreement expected by chance:

$$\mathrm{AC2} = \frac{A_o - A_e}{1 - A_e},$$

where A_o denotes the observed weighted agreement and A_e denotes the chance-expected agreement estimated under Gwet's occasional-guessing model. Since our evaluation uses 1–5 Likert-scale scores, we apply ordinal weights so that larger

score differences incur larger disagreement penalties.

For human–human reliability, AC2 is computed over all independent human ratings for each evaluated session. For human–LLM reliability, we compare the LLM-based evaluator's score with the corresponding aggregated human score for each session. Since each evaluation dimension consists of multiple Likert-scale sub-items, we first aggregate the sub-items within each dimension and then compute AC2 on the resulting dimension-level scores.

Results. As shown in Table 6, human–human AC2 ranges from 0.52 to 0.87 ($\mu = 0.72$), indicating moderate-to-strong agreement among participants. Human–LLM AC2 ranges from 0.74 to 0.85 ($\mu = 0.79$), which is comparable to human–human agreement and suggests that the LLM-based evaluator is well aligned with human judgment. Furthermore, Table 5 reports bootstrap 95% confidence intervals as error bars for the US scores in Table 3. We conduct participant-level paired analysis: for each participant, we first aggregate the US scores under IDL and PASSING then compute paired differences. We use an exact paired sign-flip test, with Holm correction across the four backbone–dataset settings. Across all four settings, PASSING achieves significantly higher US than IDL, with paired differences ranging from 1.028 to 1.280; all 95% CIs are strictly positive, and the Holm-corrected p-values are 0.007812. This confirms that PASSINGś user satisfaction is significantly higher than IDL under participant-level paired analysis.

B.9 Additional Analysis

For better understanding, we encode the five Dreyfus levels from 0 to 4 and compute mean absolute error between predicted and ground-truth levels. On MMLU-Pro and ARC-MCAS, the MAEs are 0.343 and 0.147 for GPT-4o, and 0.243 and 0.211 for DeepSeekV4-Pro, respectively. When pooling both datasets, the MAEs are 0.230 for GPT-4o and 0.224 for DeepSeekV4-Pro. Overall, 96.97% of predictions are within one level of the ground truth. We will report these ordinal metrics in the revised manuscript.

We also include the corresponding 5×5 confusion matrices in Table 7, where rows denote ground-truth Dreyfus levels and columns denote predicted levels in the order Novice, Advanced Beginner, Competent, Proficient, Expert.

C Case Study

We provide a complete case study illustrating the full PASSING pipeline: initial query → probing questions → user responses → schema filling → expertise estimation → tailored response.

Query: "Why can large-aperture telescopes see fainter objects?"

Turn 1

- What-to-Ask: S-03 (The strategy descriptions are provided at the end)
- How-to-Ask: S-14
- Inquiry: "This touches on a bit of optics — no worries if it's unfamiliar! Do you have a sense of what light-gathering power actually refers to, or how it differs from magnification?"
- User response: "I know a bigger aperture lets you see more, but I always thought that was mainly about zooming in, not really about brightness."
- Schema update: knowledge_scope → explicit; perspective → overwhelmed

Turn 2

- What-to-Ask: S-02
- How-to-Ask: S-18
- Inquiry: "Here's a comparison: if you wanted to collect more rainwater, what would you do? How might that relate to how a telescope collects light?"
- User response: "Um, maybe a wider opening catches more rain? So maybe a bigger telescope catches more light too, I guess."
- Schema update: components → context_free; action → intuitive; commitment → detached

Termination and Estimation All five schema slots are populated within 2 turns. PASSING estimates the user as Advanced Beginner.

Final Tailored Answer You're on the right track! A telescope's light-gathering power depends on the area of its aperture, not just

Setting	PASSING US (95% CI)	Difference (95% CI)	Holm p
GPT-4o / MMLU-Pro	4.580 [4.540, 4.621]	1.224 [1.073, 1.360]	0.007812
GPT-4o / ARC-MCAS	4.583 [4.573, 4.594]	1.028 [0.999, 1.055]	0.007812
DeepSeekV4-Pro / MMLU-Pro	4.423 [4.408, 4.436]	1.146 [1.056, 1.234]	0.007812
DeepSeekV4-Pro / ARC-MCAS	4.568 [4.558, 4.580]	1.280 [1.240, 1.321]	0.007812

Table 5: Error Bars and Statistical Significance in Human Evaluation.

Dataset	Backbone	Gwet's AC2 ↑		
		FA	Comp.	US
Human–Human				
ARC-MCAS	DeepSeek-V4-Pro	0.7681	0.6996	0.6641
	GPT-4o	0.6949	0.8488	0.7614
MMLU-Pro	DeepSeek-V4-Pro	0.8698	0.8121	0.7888
	GPT-4o	0.6213	0.5153	0.6477
Average		0.7385	0.7190	0.7155
Human–LLM				
ARC-MCAS	DeepSeek-V4-Pro	0.7436	0.7563	0.8096
	GPT-4o	0.7351	0.7954	0.7887
MMLU-Pro	DeepSeek-V4-Pro	0.8277	0.7479	0.7645
	GPT-4o	0.7766	0.8489	0.7419
Average		0.7708	0.7871	0.7762

Table 6: Inter-rater agreement (Gwet's AC2) for human–human and human–LLM evaluation across datasets and backbones. FA: factual accuracy; Comp.: comprehensibility; US: user satisfaction.

the width — and area scales with the square of the diameter. So doubling the aperture roughly quadruples the light collected, letting it pick up much fainter objects.

Strategy IDs used in the case example:

- S-02 (What-to-Ask): Probe its application in a different context (e.g., a practical scenario, a general rule, or a self-generated example) to elicit information regarding action, because this tests the ability to bridge declarative and procedural knowledge, distinguishing academic recall or guided analysis from a robust, applicable, and transferable understanding.
- S-03 (What-to-Ask): Probe the key distinction from a similar concept, the precise mechanism behind their intuitive language, or the nuances omitted in their simplification to elicit information regarding knowledge_scope, because this moves beyond simple recall to test the user's understanding of conceptual boundaries, surfacing foundational misunderstandings and revealing the true depth of their knowledge.
- S-14 (How-to-Ask): Probe a single, simplified component of the difficult concept — first acknowledging the difficulty to normalize it, then asking a more focused question — to elicit information regarding knowledge_scope, because it reduces cognitive load and anxiety, preventing disengagement while establishing a baseline of knowledge from which to build.
- S-18 (How-to-Ask): Probe their interpretation of or first step within a brief, concrete situation provided as a shared reference point to elicit information regarding perspective, because it lowers initial cognitive load and provides a common anchor, effectively grounding the conversation for all expertise levels and preventing novices from stalling.

To understand why IDL struggles to estimate user expertise, we compare IDL with PASSING on two representative cases with the same gold level but different predictions. As shown in Table 8, IDL makes opposite errors: it underestimates a *Competent* user in computer science when the history resembles closed-form QA, while overestimating a *Competent* user in chemistry when the history contains advanced terms such as group-14 hydrides and EPR spectroscopy. This reveals a shared failure mechanism: IDL mistakes question-level signals for user-level understanding.

These cases suggest that IDL fails because it relies on weak proxy signals rather than direct evidence of user reasoning ability. It observes historical question difficulty, domain consistency, and interaction format, but cannot observe how the user reasons through a concept. As a result, IDL infers

Table 7: Confusion matrices across different LLM backbones and datasets. Rows denote ground-truth Dreyfus levels and columns denote predicted levels in the order Novice, Advanced Beginner, Competent, Proficient, Expert.

GPT-4o / MMLU-Pro						**GPT-4o / ARC-MCAS**						**DeepSeekV4-Pro / MMLU-Pro**						**DeepSeekV4-Pro / ARC-MCAS**					
	N	AB	C	P	E		N	AB	C	P	E		N	AB	C	P	E		N	AB	C	P	E
N	**14**	0	0	0	0	N	**18**	1	0	0	0	N	**14**	0	0	0	0	N	**18**	1	0	0	0
AB	0	**12**	1	1	0	AB	2	**17**	0	0	0	AB	6	**8**	0	0	0	AB	4	**15**	0	0	0
C	0	1	**13**	0	0	C	0	1	**18**	0	0	C	0	2	**9**	3	0	C	0	3	**13**	3	0
P	0	0	0	**14**	0	P	0	0	0	**19**	0	P	0	1	0	**13**	0	P	0	0	1	**16**	2
E	0	1	5	7	**1**	E	0	0	2	6	**11**	E	0	0	0	4	**10**	E	0	0	0	6	**13**

Case	**Domain**	**Expertise**	**IDL Pred.**	**PASSING Pred.**	**IDL Error Pattern**	**PASSING Evidence**
Case 1	Computer Science	Competent	Advanced Beginner	Competent	Underestimates the user because closed-form QA history appears shallow.	Probes concept boundaries, mechanism understanding, and applied judgment.
Case 2	Chemistry	Competent	Proficient	Competent	Overestimates the user because advanced terminology appears expert-like.	Probes mechanism understanding, uncertainty, and guided reasoning.

Table 8: Case-level comparison between IDL and PASSING.

expertise from what topics the user has asked about, rather than what the user actually understands.

In contrast, PASSING actively constructs diagnostic evidence through follow-up questions. In the computer science case, PASSING probes concept boundaries, encryption mechanisms, and zero-day disclosure trade-offs. In the chemistry case, PASSING asks about M–H bond strength and further narrows the discussion to orbital overlap, revealing whether the user understands the underlying mechanism or only recognizes the general trend.

Overall, the case study shows that accurate user-level perception requires observing how users reason, not merely what topics they ask about. IDL performs proxy-based inference from historical questions, whereas PASSING performs reasoning-based diagnosis through adaptive questioning. Across cases, PASSING follows a consistent pattern: probing concept boundaries, mechanisms, uncertainty, and applied judgment. Thus, when historical signals are misleading, PASSING can still recover the correct user level. In short, IDL follows "historical question difficulty $\rightarrow$ proxy-based expertise inference", whereas PASSING follows "adaptive follow-up dialogue $\rightarrow$ reasoning-based expertise diagnosis".

D Induced Strategies

D.1 What-to-Ask Strategies

S-01 Probe their step-by-step plan or the underlying rationale for a specific choice to elicit information regarding *perspective*, because this externalizes the user's mental model, revealing their procedural thinking, priorities, and whether their actions are based on rote memorization or goal-oriented understanding.

S-02 Probe its application in a different context (e.g., a practical scenario, a general rule, or a self-generated example) to elicit information regarding *action*, because this tests the ability to bridge declarative and procedural knowledge, distinguishing academic recall or guided analysis from a robust, applicable, and transferable understanding.

S-03 Probe the key distinction from a similar concept, the precise mechanism behind their intuitive language, or the nuances omitted in their simplification to elicit information regarding *knowledge_scope*, because this moves beyond simple recall to test the user's understanding of conceptual boundaries, surfacing foundational misunderstandings and revealing the true depth of their knowledge.

S-04 Probe a boundary case, exception, ambiguous scenario, or common pitfall that challenges its standard application to elicit information regarding *perspective*, because this differentiates rigid, rule-based knowledge from a nuanced, adaptive understanding of a concept's limitations and failure modes in complex, real-world situations.

S-05 Probe the relative importance of factors, alternative approaches, trade-offs, or the limitations and critiques of their chosen approach to elicit information regarding *perspective*, because this reveals whether the user has a holistic, prioritized mental model that considers the problem space broadly, including counterarguments and limitations, rather than a linear or one-sided view.

S-06 Probe the precise, mechanistic interaction between components or other interacting systemic factors to elicit information regarding *components*, because this distinguishes static, list-based knowledge from a dynamic, systems-level understanding by forcing the user to trace causal links and consider the system holistically.

S-07 Probe the informal signals, 'gut feelings,' or intuitive patterns they use to guide analysis or form hypotheses to elicit information regarding *action*, because this differentiates a competent user's deliberate process from an expert's intuitive pattern-matching, revealing the shift from analytical to holistic decision-making.

S-08 Probe the specific source of their hesitation or their method for reconciling the inconsistency to elicit information regarding *commitment*, because this forces the user to articulate the precise boundaries of their own knowledge, revealing how their knowledge is structured and whether they can resolve internal conflicts.

S-09 Probe their personal experiences, professional stance, or the most difficult 'real-world' aspect of the issue to elicit information regarding *commitment*, because an expert's deep experience is often linked to a strong sense of personal or professional investment, and eliciting this can help differentiate a detached analyst from an involved expert.

S-10 Probe the underlying first principles, policy reasons, or historical context for that rule to elicit information regarding *knowledge_scope*, because this tests for tacit understanding of a rule's origin and purpose ('the why') rather than just explicit knowledge of its content ('the what'), a key differentiator between Competent practitioners and Proficient or Expert thinkers.

S-11 Probe the limitations, critiques, or weaknesses of that same concept or model to elicit information regarding *perspective*, because a Competent user can accurately apply a model, but an Expert understands its boundaries and when it breaks down. This probe forces a shift from a deliberate explanation of the model's parts to a holistic evaluation of the model itself.

S-12 Probe a counter-intuitive scenario where that causal link is broken or reversed to elicit information regarding *action*, because analytic reasoning follows predictable causal chains, which is characteristic of Competence. Asking for exceptions or paradoxes tests for an intuitive grasp of the topic, which allows for non-linear thinking and is a hallmark of an Expert.

S-13 Probe a request for synthesis and prioritization across the discussed topics (e.g., 'What is the single most overlooked factor?') to elicit information regarding *perspective*, because such signals suggest a user may have a deeper, more integrated understanding than they have explicitly stated. Asking them to synthesize and prioritize forces them to move beyond explaining individual components and demonstrate a holistic view of the entire system, revealing Proficient or Expert-level thinking.

D.2 How-to-Ask Strategies

S-14 Probe a single, simplified component of the difficult concept — first acknowledging the difficulty to normalize it, then asking a more focused question — to elicit information regarding *knowledge_scope*, because it reduces cognitive load and anxiety, preventing disengagement while establishing a baseline of knowledge from which to build.

S-15 Probe the consequence or increasing complexity of their correct foundational answer

— building on it affirmingly to incrementally raise the scope (e.g., 'Exactly. Building on that, what is the consequence of...?') — to elicit information regarding *perspective*, because it creates a logical and encouraging conversational flow, confirming their base knowledge while smoothly gauging the depth of their understanding.

S-16 Probe the 'why' behind their 'what' — the underlying principle or reasoning they have not yet justified (e.g., 'Can you walk me through your thinking?' or 'What's the underlying principle for that step?') — to elicit information regarding *perspective*, because it elicits the user's mental model, distinguishing deep, principle-based understanding from rote memorization.

S-17 Probe the precise distinction between two closely related concepts — whether the user has defined one correctly or only partially — (e.g., 'What's the primary difference between X and Y?') to elicit information regarding *knowledge_scope*, because it tests the precision and boundaries of their understanding, revealing whether their knowledge is superficial or nuanced and interconnected.

S-18 Probe their interpretation of or first step within a brief, concrete situation provided as a shared reference point to elicit information regarding *perspective*, because it lowers initial cognitive load and provides a common anchor, effectively grounding the conversation for all expertise levels and preventing novices from stalling.

S-19 Probe a concrete, self-generated example from their own experience that grounds the abstract concept they described (e.g., 'Can you give me a specific example of when you've seen that happen?') to elicit information regarding *components*, because it tests for active recall and grounds abstract knowledge in practical application, differentiating between theoretical knowledge and applied expertise.

S-20 Probe how their approach would shift when a trade-off, conflicting goal, or ambiguous constraint is introduced into the scenario (e.g., 'How would you prioritize if you had limited resources?' or 'What if constraint X were introduced?') to elicit information regarding *action*, because it tests the robustness and flexibility of a user's mental model, revealing their ability to adapt and make judgments under pressure — a hallmark of true expertise.

S-21 Probe how a different stakeholder would view or evaluate the same concept or decision (e.g., 'How would the finance team view this decision differently?') to elicit information regarding *perspective*, because it probes for a systems-level understanding and empathy, which are hallmarks of higher proficiency, moving beyond self-contained knowledge.

E Prompts

The knowledge graph construction prompt
Input: <Query>, <Answer>
Output 1: <Knowledge Dependency Graph with min-level> (for user simulator)
Output 2: <Knowledge Dependency Graph without min-level> (for probing)

Table 9: The KG construction. The *min-level* attribute is masked from the probing agent to prevent information leakage.

The What-to-Ask strategy selection prompt
Your task is to decide what information to ask about, using the *What-to-Ask* strategy set.
Inputs:
<Conversation History>
<Populated Schema>
<*What-to-Ask* Strategy Set>
Output: <Information to Ask>

Table 10: The *What-to-Ask* prompt for deciding what information to ask about.

The How-to-Ask strategy selection prompt
Your task is to decide how to ask that question, using the *How-to-Ask* strategy set.
Inputs:
<Information to Ask>
<Conversation History>
<*How-to-Ask* Strategy Set>
Output: <Formulated Question>

Table 11: The *How-to-Ask* prompt for deciding how to ask that question.

The inquiry generation prompt

Your task is to generate the inquiry to ask the user.
Inputs:
<Formulated Question>
<Conversation History>
Output: <Inquiry>

Table 12: The inquiry generation prompt.

The user simulator prompt

Your task is to roleplay as a real person and respond to the inquiry, strictly following the persona, expertise level, and knowledge boundaries specified in the inputs below.
Inputs:
<Persona Description>
<Dreyfus Level>, <Known Concepts>, <Unknown Concepts>
<Refusal Strategies>
<User Query>, <Conversation History>, <Inquiry>
Output: <User Response>

Table 13: The user simulator prompt.

The slot filling prompt

Your task is to extract and fill information into the schema based on the user response and conversation history.
Inputs:
<User Response>
<Conversation History>
Output: <Updated Schema with Filled Slots>

Table 14: The slot filling prompt.

The expertise estimation prompt

Your task is to estimate the user's expertise level based on the collected information.
Inputs:
<Filled Schema>
<Conversation History>
Output: <Estimated Expertise Level>

Table 15: The expertise estimation prompt.

The response generation prompt

Your task is to provide the final answer to the user's initial question.
Inputs:
<Filled Schema>
<Conversation History>
<Estimated Expertise Level>
Output: <Final Answer>

Table 16: The response generation prompt.