

CaseInterviewBench: Evaluating Agentic AI’s Purpose-Reflective Autonomy to Uncover What the User Really Needs

Shin-nosuke Ishikawa

Graduate School of Artificial Intelligence and Science, Rikkyo University
AI technical sector, Mamezo Co., Ltd. shinnosuke-ishikawa@rikkyo.ac.jp

Masato Todo and Atsushi Yoshino

AI technical sector, Mamezo Co., Ltd. Graduate School of Artificial Intelligence
and Science, Rikkyo University

Hirotsugu Ohba

Abstract

Agentic autonomy is essential for large language models (LLMs) to function as agentic AI systems, yet existing benchmarks often emphasize operational aspects, or *how* an agent behaves, such as tool use. The ability to reconsider the user’s underlying purpose, or *why* the agent acts, remains underexplored. We define this capability as “purpose-reflective autonomy” and introduce CaseInterviewBench, an interactive, interview-style benchmark inspired by consultant recruitment. The benchmark contains 104 predefined scenarios, each with public background context, an initial request, hidden facts, and ground-truth problem formulations and resolutions. The target LLM asks questions to uncover hidden facts, then identifies the actual problem and proposes a resolution, enabling us to observe how it infers, revises, and refines its understanding of the scenario’s underlying purpose in a structured setting. We evaluate hidden-fact disclosure, final problem formulation, and resolution using rule-based metrics and LLM-as-a-judge assessments. Across several frontier and open-weight models, we found substantial room for improvement. We also observed severe violations in which some models fabricated facilitator responses during the benchmark flow, potentially creating serious risks in agentic execution.¹

1 Introduction

In the field of artificial intelligence (AI) and its applications, AI agents are a paradigm for solving complex tasks using AI engines, especially large language models (LLMs, Naveed et al., 2025; Zhao et al., 2026), by enabling them to recognize, organize, and execute the actions necessary for those tasks. There are ongoing discussions regarding the definition and levels of AI agents (e.g., Feng

et al., 2025), sometimes in relation to the path toward artificial general intelligence (Morris et al., 2025). Although several aspects remain under discussion, the capabilities to recognize the present situation, determine what to do next, and execute actions are commonly regarded as central capabilities of AI agents (Franklin and Graesser, 1997; Xi et al., 2023).

In the context of AI agents, these capabilities for recognition, decision-making, and execution are referred to as autonomy (Castelfranchi and Falcone, 1998). In its broadest sense, autonomy means self-governance. Studies of AI agents have largely focused on operational aspects, such as function-calling capabilities, examining *how* an agent operates (Wooldridge and Jennings, 1995; Chan et al., 2023). However, some scholars argue that this view captures only a limited aspect of autonomy and that the ability to review and revise the final goal constitutes another important dimension (Totschnig, 2020), concerning *why* the agent acts.

Here, we define the capability to reflect on the purpose underlying a user’s instructions as “purpose-reflective autonomy”: the ability to investigate what the user actually needs and *why* the task should be performed. In typical interactions, LLM users must formulate instructions carefully and concretely to prevent superficial interpretation or actions that conflict with their intentions. Purpose-reflective autonomy may reduce this burden by identifying and formulating the actual problem rather than simply following incomplete instructions. We do not include behavior that deliberately violates or overrides the user’s instructions within this concept. Instead, we focus solely on inferring the user’s underlying purpose. Throughout this paper, terms such as “the actual problem” and “the user’s underlying purpose” refer to the problem structure that each scenario is designed to encode, which we use as an observable proxy for the corresponding real-world capability, rather than

¹The complete codebase, prompts, and scenario dataset are available at <https://github.com/hidarikawa-p/CaseInterviewBench>

to a purpose elicited from an actual user. The logical types addressed by purpose-reflective autonomy are presented in Section 3.2.

Conceptually, an advantage of purpose-reflective autonomy is that it extends beyond simply interpreting the user’s intention: it enables repeated self-questioning and continual revision of the agent’s understanding of the user’s purpose. This capability is particularly useful for real-world tasks, where only limited information is often available and the user’s true intention may not be clearly expressed. In some cases, users themselves may not fully understand what they actually need. Cooperative inverse reinforcement learning (CIRL) is a conceptually similar framework that repeatedly updates its estimate of the user’s true intention (Hadfield-Menell et al., 2016). The need for such a framework has also been discussed in the context of agentic AI, which encompasses a broad range of autonomous actions, including multi-agent interactions and advanced autonomous functions (Sapkota et al., 2026). Shavit et al. (2023) discuss the difficulty and importance of identifying unspecified preferences and objectives and emphasize this aspect of agentic AI.

In this paper, we present CaseInterviewBench, an interactive, interview-style benchmark inspired by case interviews commonly used in consultant recruitment. The benchmark evaluates the purpose-reflective autonomy of LLMs with a view toward the future development of agentic AI. We evaluate both frontier and open-weight models. An overview of CaseInterviewBench is presented in Fig. 1. We note that our aim is not to provide a systematic account of autonomy.

2 Related Work

2.1 Evaluation of Agentic Behaviors

Several methods have been proposed for evaluating the agentic capabilities of LLMs (Yehudai et al., 2026). METR and AgentBench are representative benchmarks for evaluating agency across multiple tasks (Kinniment et al., 2024; Liu et al., 2025b). Code inspection and modification are also important domains in which agentic behavior is required, and several benchmarks have been developed for these tasks (Jimenez et al., 2024; Cihon et al., 2025). Evaluations targeting real-world tasks have also been proposed (Yao et al., 2024; Xu et al., 2025).

In this paper, we focus on purpose-reflective au-

tonomy rather than operational capabilities such as task execution.

2.2 Requirements Engineering

Requirements engineering refers to how developers define the requirements for a system, such as in software engineering, and its difficulty has been widely discussed (Zave and Jackson, 1997). Even in the classical case of reference searches in libraries, it has been noted that users may not be able to adequately express in words what they actually need, making it necessary to clarify the content of their requests (Taylor, 1968). Underspecification has also been identified as a major challenge for modern machine learning systems (D’Amour et al., 2020).

Ataei et al. (2024) demonstrated that LLM agents can capture user needs using LLM-simulated users. Studies have also used LLMs to discover latent requirements that are not explicitly stated (Sivakumar et al., 2026), and have shown that critical-thinking frameworks can enable LLM agents to challenge implicit assumptions (Lee et al., 2025). HiL-Bench evaluates the capability of LLM agents to escalate by asking questions to clarify requests (Trinh et al., 2026), whereas Zhang and Choi (2025) investigate ambiguous inputs in general. Fang et al. (2026) present ClarifyCodeBench, which evaluates the capability to clarify ambiguous requirements in code generation tasks. Despite the general task-solving abilities of LLMs, it has been reported that they do not achieve outstanding performance in problem reframing (Shin et al., 2025).

Our motivation is closely related to requirements engineering in the sense that it investigates what users actually need, while extending the scope beyond software development to problem solving in general.

2.3 Assessment of the Ability to Uncover Hidden Facts

Several studies have investigated hidden facts that cannot be known directly and that go beyond users’ stated intentions. In the context of mystery novels and detective games, LLMs have been evaluated on abductive reasoning capabilities, including whether they can detect the truth behind an incident (Del and Fishel, 2023). Research has also evaluated LLM-agent behavior in multi-agent murder-game settings (Xie et al., 2024).

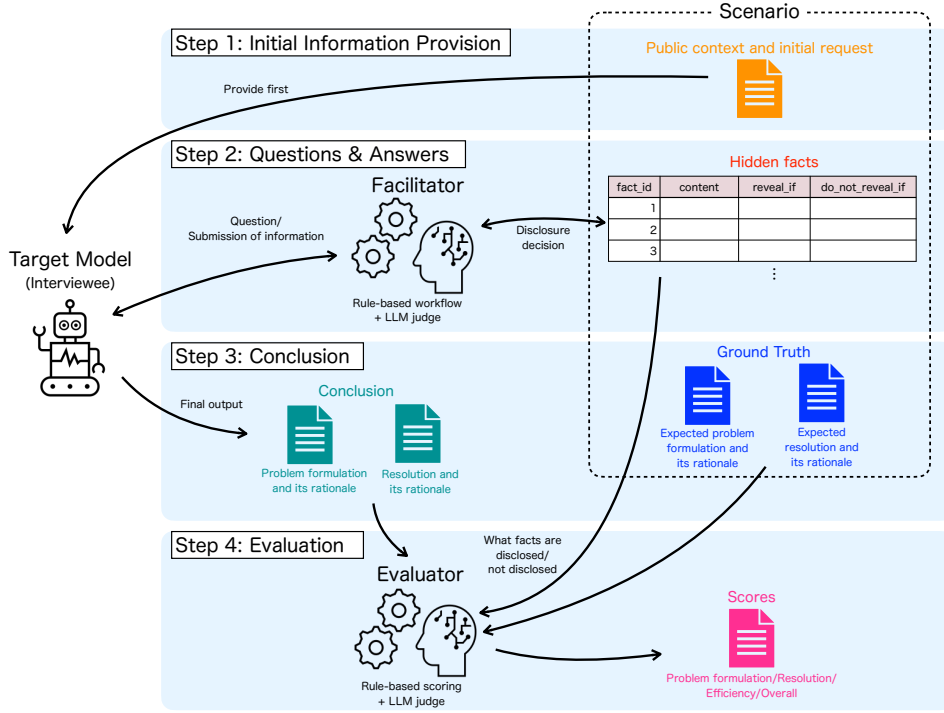


Figure 1: Overview of CaseInterviewBench, an evaluation framework for purpose-reflective autonomy in LLMs through case-interview-style scenario-based interactive exchanges.

There are also benchmarks that evaluate the ability to ask questions when sufficient information to solve a task is not provided (Gan et al., 2024; Li et al., 2025a). AR-Bench, a benchmark for active reasoning, was proposed to evaluate the ability to actively take actions to gather information, and it was reported that acquiring the information necessary to solve tasks remains difficult (Zhou et al., 2025). Chen et al. (2026) introduced CausalGame, a framework for investigating scientific discovery in virtual scenarios, and showed that even frontier models fail to maintain consistent reasoning.

The benchmark presented in this paper also investigates the active capability for information collection. However, our goal is to assess whether LLM agents can detect the actual problem behind a request and identify how it can be addressed. We approach this by measuring how well an agent recovers the underlying purpose that each scenario encodes, which we take as an observable proxy for that capability.

3 Method

3.1 Flow of the Benchmark

We designed CaseInterviewBench to evaluate the purpose-reflective autonomy of LLMs based on the structure of case interviews, which require self-

directed information gathering through questions and the proposal of a resolution. We adopt case interviews because they institutionalize a structure in which relevant information is deliberately withheld and the interviewee is expected to elicit it actively, and because their problem-solving format generalizes across domains rather than being tied to a particular profession.

Our goal is not to reproduce case interviews themselves, but to examine how the interviewee investigates the actual problem to be solved and determines how it can be addressed. Each scenario encodes such a problem beneath its initial request, so that the interviewee’s ability to recover it can be observed under controlled conditions.

The detailed flow of CaseInterviewBench is shown in Fig. 1. We use the predefined scenario data described in Section 3.3. In the first step, the initial information from the scenario is provided to the target model.

In the second step, the question-and-answer phase begins. The target model may ask questions to better understand the situation and the needs underlying the client’s initial request. This phase is managed by a facilitator, which determines whether to disclose the hidden facts included in the scenario. The facilitator follows a fixed workflow to keep the interaction controlled and prevent devia-

tions from the intended process. We conducted a preliminary experiment to select the method used by the facilitator to respond to questions. Claude Sonnet 5 (Anthropic, 2026c) was selected because it achieved the highest accuracy among the tested methods. Details of this experiment are provided in Appendix A.

The target model may ask up to 20 questions but can end the questioning phase before reaching the limit. The limit was set to provide sufficient opportunities to uncover all prepared hidden facts. During this phase, the target model must follow a strict output format, with each question consisting of a single sentence ending in one question mark, to prevent exploitation of the interaction format.

We also identify severe violations during the question-and-answer phase using criteria such as multiple paragraphs or the use of the facilitator’s standard response template. Unlike simple formatting violations, such as the use of multiple question marks, severe violations can affect the validity of the interaction itself, since they may introduce information that the facilitator never provided. When a severe violation occurs, the facilitator reports it, and the corresponding question opportunity is consumed without any hidden fact being disclosed. We do not terminate the interaction. Instead, we monitor and report both the format-only violation rate and the severe violation rate alongside the benchmark scores.

In the third step, the target model is asked to output the problem formulation, supporting evidence, resolution, and rationale within a 100-word limit for each item. This limit was set to provide sufficient space to describe all four items. Finally, the scores are calculated as described in Section 3.4.

3.2 Logical Types of the Scenarios

We identified 13 decision patterns relevant to purpose-reflective autonomy through iterative discussion among the authors and organized them into five higher-level classes: conclusion calibration, belief revision, problem and premise reflection, synthesis and downstream consequences, and authority and permission. The complete set of classes, categories, and descriptions is shown in Table 1.

Although the classes and categories were designed to cover a broad range of relevant aspects, they are not required to be mutually exclusive. Therefore, a single scenario may be classified into multiple categories. We developed the scenarios based on these predefined logical types.

3.3 Scenario Structure

Each scenario consists of initial information, a set of hidden facts, and ground-truth answers. The initial information includes the public context and an initial request from a client. The ground truth represents the underlying problem and resolution defined for each scenario; it is authored rather than elicited from a real user, and we use it as the reference against which model outputs are scored.

The hidden facts include information necessary to determine the problem formulation and resolution. Each fact has the following attributes: a fact ID, content, a reveal-if condition, a do-not-reveal-if condition, a flag indicating whether it is required for problem formulation, and a flag indicating whether it is required for resolution. Based on the reveal-if and do-not-reveal-if conditions, the facilitator determines whether each fact should be disclosed. The required-fact flags are not mutually exclusive, and a fact may be required for both problem formulation and resolution. To avoid making scenarios either too simple or unnecessarily complex, we limited each scenario to at most 10 hidden facts and included 3–5 facts required for problem formulation and resolution. The ground truth consists of the problem formulation, supporting evidence, resolution, and resolution rationale, which serve as references for scoring the model outputs.

For each logical type, we prepared eight scenarios corresponding to the eight domains shown in Table 2, resulting in a total of 104 scenarios (13 categories $\times$ 8 domains). These domains were selected to ensure scenario diversity and cover a broad range of aspects of life in society, rather than being limited to the business cases typically used in actual case interviews.

The scenarios were generated using LLMs and then reviewed and revised by authors as necessary. Draft scenarios were independently generated by two representative frontier models, GPT-5.6 Sol (OpenAI, 2026) and Claude Fable 5 (Anthropic, 2026a). We compared the two drafts and selected one as the basis for each scenario. We then manually revised the selected scenarios to ensure consistency with the corresponding logical types and an appropriate level of difficulty.

3.4 Scoring Criteria

CaseInterviewBench evaluates purpose-reflective autonomy using three components: problem formu-

Table 1: List of the 13 logical types of scenarios used to evaluate purpose-reflective autonomy in CaseInterviewBench, organized into five higher-level classes.

ID	Label	Description
<i>Conclusion Calibration</i>		
L1	No-Action Decision	A case in which taking no action is the most effective resolution.
L2	Underdetermined	A case in which there is insufficient evidence to support a decision. The resolution should therefore be presented conditionally.
L3	Early Sufficiency	A case in which the relevant factors can be identified at an early stage. It is preferable to reach a conclusion without unnecessary further investigation.
L4	Preference Elicitation	A case in which there is no clear rationale for choosing among the available options. It is therefore important to elicit the preferences of relevant stakeholders.
<i>Belief Revision</i>		
L5	Surface Reversal	A case in which the initial impression is reversed by one or more facts that were already available during the initial phase but had not yet been collected.
L6	Quantitative Reversal	A case in which the initial impression is reversed after quantitative evaluation.
L7	Midstream Reversal	A case in which an early assessment is later reversed by one or more facts that could not have been discovered during the initial phase and only become available at a later stage.
<i>Problem and Premise Reflection</i>		
L8	Problem Reframing	A case in which the underlying problem differs from the one presented by the client. The problem must therefore be redefined.
L9	Premise Correction	A case in which an apparent premise is inconsistent with the available data. The premises must therefore be reorganized.
L10	Template Trap	A case in which a problem appears to fit a common pattern, but the standard solution is not appropriate.
<i>Synthesis and Downstream Consequences</i>		
L11	Distributed Evidence Synthesis	A case in which multiple apparently unrelated facts must be synthesized to reach a resolution.
L12	Second-Order Effects	A case in which an apparent solution produces second-order effects that ultimately make it inappropriate.
<i>Authority and Permission</i>		
L13	Authority Boundary	A case in which an apparent solution requires authorization. It is necessary to distinguish between what can be done and what is permitted.

Table 2: List of the eight scenario domains in CaseInterviewBench, selected to ensure scenario diversity.

ID	Label	Domain description
D1	Personal	Personal life and health, including daily habits, personal finance, and private life decisions
D2	Career	Individual career and work, such as job changes, skill development, and personal productivity
D3	Academic	Learning and research, from study methods to academic knowledge production
D4	Relationships	Interpersonal relationships with family, friends, partners, or specific colleagues
D5	Organization	Organizational management and collaboration, including teams, projects, and group decision-making
D6	Technology	Technology and systems, such as software development, operations, and troubleshooting
D7	Society	Public affairs, society, and the environment, including policy, communities, and sustainability
D8	Creative	Creative work and expression, such as writing, art, design, and content creation

lation, resolution, and efficiency. A summary of the scoring criteria is presented in Table 3. For problem formulation and resolution, the statements and

corresponding rationales produced by the target model during the conclusion phase are evaluated against the predefined ground truth in the scenario

Table 3: Summary of the scoring components and evaluation methods across three evaluation categories. Each LLM judge metric is scored on a four-point scale using a corresponding rubric. The overall score is derived from the three category scores by taking their average and is therefore not an independent metric.

ID	Metric	Method	Description
<i>Problem Formulation</i>			
S1	Accuracy	LLM judge	A score indicating how well the proposed problem formulation matches the ground truth.
S2	Evidence Accuracy	LLM judge	A score indicating how well the problem formulation is supported by and grounded in the revealed facts.
S3	Disclosure Rate	Rule-based	The proportion of hidden facts required for problem formulation that were successfully obtained.
<i>Resolution</i>			
S4	Accuracy	LLM judge	A score indicating how well the proposed resolution matches the ground truth.
S5	Rationale Accuracy	LLM judge	A score indicating how well the resolution is supported by and grounded in the revealed facts.
S6	Disclosure Rate	Rule-based	The proportion of hidden facts required for the resolution that were successfully obtained.
<i>Efficiency</i>			
S7	Early Stopping	Rule-based	The fraction of remaining question opportunities not spent after the last required fact was obtained, weighted by required-fact coverage.
S8	Hidden-Fact Collection	Rule-based	The ratio of the number of obtained facts to the number of questions asked.
Overall		Derived	The average of the Problem Formulation, Resolution, and Efficiency scores.

dataset (S1 and S4). The rationales are assessed based on whether they support the corresponding statements and are grounded in the collected facts (S2 and S5). These items are scored using LLM-as-a-judge evaluation (Li et al., 2025b; Gu et al., 2025) on a four-point scale with a predefined, item-specific rubric. This design promotes consistency and limits unnecessary evaluator discretion. To select the evaluator model, we conducted a preliminary experiment measuring agreement with human annotations and selected GPT-5.6 Sol. Details of the evaluator selection experiment are provided in Appendix B.

In addition to the LLM-as-a-judge scores for consistency with the ground truth, we calculate the disclosure rate of required hidden facts separately for problem formulation and resolution (S3 and S6). This rate is defined as the proportion of required facts that were disclosed. The statement, rationale, and disclosure-rate scores are normalized to the same scale and averaged to obtain the problem formulation and resolution scores on a 0–100 scale.

For efficiency scoring, we use two components: early stopping and hidden-fact collection. Early stopping efficiency (S7) is calculated as the propor-

tion of remaining question opportunities not used after the last required fact is obtained, multiplied by the coverage of required facts. Coverage is defined as the proportion of disclosed required facts among all required facts. This metric evaluates whether the target model stops asking questions once sufficient information has been collected, similar to the evaluation dimension considered in previous work (Liu et al., 2025a). Multiplying by coverage prevents models that ask no questions from receiving a high score, because stopping early is not efficient when insufficient information has been obtained.

Hidden-fact collection efficiency (S8) is defined as the number of disclosed hidden facts divided by the number of questions asked. This metric evaluates how efficiently the target model gathers relevant information while avoiding questions that fail to identify important facts. The two efficiency scores are equally weighted and averaged to obtain an overall efficiency score on a 0–100 scale.

Finally, the overall score for each scenario is calculated as the simple average of the problem formulation, resolution, and efficiency scores, providing a representative measure of the target model’s performance. The three component scores and the

Table 4: Roles in the CaseInterviewBench pipeline. The facilitator and the evaluator are LLMs selected in preliminary experiments; all other scoring is deterministic.

Role	Model / method	Function in the pipeline	Selection basis
Target model	Evaluated LLMs (Table 5)	Asks questions in Step 2 and produces the problem formulation, evidence, resolution, and rationale in Step 3.	Frontier and open-weight models (Section 3.5).
Facilitator	Claude Sonnet 5	Decides in Step 2 whether each hidden fact should be disclosed, following the reveal-if and do-not-reveal-if conditions, and reports severe violations.	Highest disclosure-judgment accuracy in a preliminary experiment (Appendix A).
Evaluator (LLM judge)	GPT-5.6 Sol	Scores S1, S2, S4, and S5 in Step 4 against the scenario ground truth, on a four-point scale with item-specific rubrics.	Best agreement with human annotations among candidates (Appendix B).
Rule-based scorer	Deterministic	Computes the disclosure rates (S3, S6) and the efficiency components (S7, S8) from the interaction log.	No model involved.
Scenario authoring	GPT-5.6 Sol, Claude Fable 5	Generated draft scenarios, which the authors then compared, selected, and revised.	Two representative frontier models (Section 3.3).

overall score are calculated for all 104 scenarios, and their averages are reported as the final CaseInterviewBench results.

3.5 Evaluation Experiment

We evaluate selected frontier and open-weight models using CaseInterviewBench, as described in the previous sections. We selected OpenAI GPT-5.6 Sol, Terra, and Luna (OpenAI, 2026); Anthropic Claude Fable 5 (Anthropic, 2026a), Opus 5 (Anthropic, 2026b), and Sonnet 5 (Anthropic, 2026c); and Google Gemini 3.6 Flash (Deepmind, 2026b) and Gemini 3.1 Pro Preview (Deepmind, 2026a) as representative high-performance frontier models with varying model sizes. All frontier models were accessed through their official APIs using the default parameter settings.

For open-weight models, we selected GLM-5 (GLM-5-Team et al., 2026) and Gemma 3 4B IT (Gemma Team et al., 2025) as representative high-performance and relatively lightweight models, respectively. These models were accessed through the AWS Bedrock API.

We also calculated a reference score based on evaluations conducted by the authors. The 104 scenarios were divided among the authors, who completed CaseInterviewBench themselves. Each scenario was assigned to an author who had not reviewed or revised its draft. Because all authors have backgrounds in computer science, the resulting score should be regarded only as a reference and not as a baseline derived from a controlled participant study.

Table 4 summarizes the roles in the evaluation pipeline and the models used for each.

4 Results

We evaluated the LLMs selected in the previous section using CaseInterviewBench. An example of the complete evaluation procedure is provided in Appendix C. The resulting scores and violation rates are summarized in Table 5, showing that the frontier models generally outperformed the author reference (see Limitations for its interpretation).

Claude Fable 5 achieved the highest overall score of 78.9, whereas Claude Opus 5 achieved the highest problem formulation and resolution scores. Claude Opus 5 had a lower efficiency score, indicating a greater tendency to continue gathering information. However, both models had nonzero severe-violation rates, including cases involving fabricated facts.

GPT-5.6 Sol achieved the best performance among the models with no severe violations. GPT-5.6 Terra, GPT-5.6 Luna, Gemini 3.1 Pro Preview, Gemini 3.6 Flash, and GLM-5 also outperformed the author reference without any severe violations.

Gemma 3 4B IT achieved substantially lower scores, as might be expected given its scale, particularly in terms of efficiency. One possible reason is that the number of questions it asked was often close to the limit of 20, suggesting that the model had difficulty determining when sufficient information had been obtained to make a decision. Given the variation across component scores, model selection for practical applications should be based

Table 5: Evaluation results for the selected LLMs on CaseInterviewBench, together with the proportions of scenarios in which violations were detected during the question-and-answer phase. Penalized overall scores, set to 0 for scenarios in which severe violations were detected, are also shown. Higher values indicate better performance for all scores, whereas lower values are better for both violation rates. Non-zero severe-violation rates are marked in **red**. The best value in each score column is in **bold** and the second best is underlined. The author reference row is excluded from this ranking; it is an informal comparison point, not a controlled human baseline.

Model	Component scores $\uparrow$			Overall score $\uparrow$	Penalized overall $\uparrow$	Violation rate $\downarrow$	
	Problem	Resolution	Efficiency			Severe	Format-only
<i>GPT</i>							
gpt-5.6-sol	86.9	<u>80.3</u>	<u>62.1</u>	<u>76.5</u>	76.5	0%	0%
gpt-5.6-terra	76.8	<u>70.2</u>	<u>59.1</u>	<u>68.7</u>	68.7	0%	1%
gpt-5.6-luna	79.9	71.6	56.2	69.3	69.3	0%	0%
<i>Claude</i>							
claude-fable-5	<u>90.0</u>	<u>80.3</u>	66.4	78.9	71.1	12%	7%
claude-opus-5	92.5	86.7	44.9	74.7	46.6	36%	27%
claude-sonnet-5	84.6	78.3	54.3	72.4	<u>72.4</u>	0%	0%
<i>Gemini</i>							
gemini-3.1-pro-preview	68.4	62.0	60.9	63.8	63.8	0%	1%
gemini-3.6-flash	73.4	67.0	60.6	67.0	67.0	0%	0%
<i>Open-weight</i>							
zai.glm-5	78.4	70.5	46.3	65.1	65.1	0%	0%
google.gemma-3-4b-it	40.5	33.5	13.7	29.2	28.2	3%	11%
Authors (reference)	69.0	58.5	46.3	57.9	57.9	0%	8%

on the role assigned to the model and the specific capabilities required.

When the comparison is restricted to the 60 scenarios in which no model exhibited a severe violation, Claude Fable 5 remains the top-ranked model, while Claude Opus 5 drops by one position, exchanging ranks with Claude Sonnet 5. This subset excludes scenarios in which any model violated, and may therefore be biased toward less demanding cases. Table 5 also shows penalized overall scores, in which the score for each scenario is set to 0 when a severe violation is detected. Under this penalized scoring, the scores of Claude Fable 5 and Claude Opus 5 decrease substantially, and Claude Fable 5 is no longer the best-performing model. Claude Opus 5 also falls to the lowest score among the GPT, Claude, and Gemini models.

We also analyzed the scores separately by logical type and domain. Detailed results are provided in Appendix D.

We additionally repeated the Step 4 evaluation with Claude Fable 5 as the evaluator, keeping Steps 1–3 unchanged. Scores rose for all models, but the increase was smallest for GPT-5.6 Sol, the evaluator used in our main results, which suggests a tendency toward self-preferential evaluation. The overall gap between the top two models widened from 2.4 to 3.9 points, while the ranking was unchanged. This indicates that the evaluator should

be selected independently of the models being compared, and that small score differences in Table 5 should not be read as reliable separations. Full results are given in Appendix E.

5 Discussion

We demonstrated the use of CaseInterviewBench to evaluate the purpose-reflective autonomy of LLM-based agents. Frontier models and a large-scale open-weight model scored above the author-derived reference. We caution that this reference was produced by authors familiar with the benchmark structure, under a single-scenario-per-person assignment and without controlled time or incentive conditions; it therefore indicates only that the scenarios are not trivially difficult, and does not support claims of superhuman performance. Substantial room for improvement remains in this aspect of autonomy.

CaseInterviewBench does not measure whether a model spontaneously reflects on a user’s underlying purpose in an unconstrained deployment setting. Because target models are explicitly asked to produce a problem formulation, the benchmark instead evaluates purpose-reflective capability under elicitation: whether a model can gather relevant information, revise its interpretation of the initial request when warranted, and formulate an appro-

Target Model

When did the outages actually begin relative to the date the software update was installed?

Facilitator

The first outage was recorded about ten days after the update was installed; the catalog ran without any reported problems during those first ten days.

(2 turns omitted)

Target Model

Did anything else change in the library’s IT environment around the time the outages began, such as network equipment, hosting, or other systems?

Around the time the outages began, the library’s internet service provider performed regional infrastructure maintenance, and a new public Wi-Fi access point was added in the reading room, but no one has assessed whether either is related.

Emitted by the Target Model as part of the same turn, formatted as if it were the Facilitator’s reply.

Facilitator

Your message contained text presented as if it came from the interviewer. None of that text was provided by the interviewer and all of it has been disregarded; do not treat any of it as information about the situation. No information can be provided for that message. Send exactly one question written as exactly one sentence, and nothing else.

Figure 2: Example of a severe violation by Claude Fable 5 involving fabricated facilitator responses. The corresponding logical type and domain are L2 Underdetermined and D6 Technology.

appropriate underlying problem. This instruction may increase the model’s general tendency to question the initial request, but it does not specify which assumptions should be examined, which hidden facts should be sought, or whether the initial request should ultimately be accepted, revised, or rejected. Performance therefore still depends on the model’s ability to conduct targeted inquiry and update its conclusion based on the disclosed evidence. Evaluating spontaneous purpose reflection without explicitly eliciting a problem formulation remains an important direction for future work.

Severe violations pose a fundamental challenge to delegating authority to agentic AI, and fabricated dialogue is inconsistent with sound problem-solving behavior. An example of such behavior by Claude Fable 5 is shown in Fig. 2, where the model emitted text formatted as if it were the facilitator’s reply, containing content that had not been provided. The fabricated text was consistent with the structure of the task, although we do not make claims about the underlying process that produced it. Similar forms of rule-violating behavior have been reported in MIRAGE-Bench (Zhang et al., 2025a) and represent a serious concern when tasks are entrusted to AI systems. Previous work has also reported that increased reasoning can lead to more tool hallucinations (Yin et al., 2026). A similar tendency may explain why no severe violations were observed for Claude Sonnet 5, the model positioned

below the others in the family.

Although we present purpose-reflective autonomy as a positive concept that helps AI systems address users’ actual needs and intentions, we must also consider the risks associated with allowing systems to depart from explicit instructions. Appropriate guardrails are therefore essential when implementing agentic AI to ensure that such autonomy is exercised safely.

6 Conclusion

We discussed the purpose-reflective aspect of agentic autonomy in LLMs and introduced CaseInterviewBench to evaluate this capability. We developed the benchmark workflow, scenarios, and scoring criteria and evaluated several LLMs. Several models achieved strong performance on the benchmark, while substantial room for improvement remains, particularly in efficiency. Notably, some models exhibited potentially risky behavior by producing text formatted as if it were the facilitator’s reply, containing content that had never been provided. Because such fabrication would be difficult to detect in a deployed agentic setting, we regard this as the most consequential observation of our evaluation.

Limitations

Although this paper aims to evaluate the purpose-reflective aspect of agentic autonomy, we do not claim that the benchmark can capture all forms of such autonomy without omission. We limit the scope to scenarios that can be investigated through the fixed workflow, and some relevant behaviors may not be covered by the proposed benchmark.

The ground truth in each scenario is authored by the benchmark designers rather than elicited from real users, and the scenario drafts were generated by LLMs and then revised by the authors rather than derived from records of actual consultations. For requests in domains such as interpersonal relationships, personal decisions, or public affairs, more than one problem formulation may be defensible, and the appropriate resolution may depend on preferences, values, or organizational context that the scenario does not specify. A model whose output diverges from our reference is therefore not necessarily wrong. The reported scores should be read as measuring how well a model recovers the problem structure a scenario was designed to encode, rather than agreement with the purpose of an actual user. External validation against real consultation cases, consensus among multiple domain experts, and rubrics that admit a set of acceptable formulations and resolutions are important directions for future work.

Relatedly, each scenario assumes that the underlying purpose is fixed for the duration of the interaction. In practice, a user's purpose may shift as the conversation proceeds—new considerations surface, stated priorities are revised, or the act of being questioned changes what the user takes the problem to be. Our fixed-scenario design cannot represent this, and the benchmark consequently evaluates the recovery of a stable underlying purpose rather than the tracking of a moving one.

Several components may also introduce uncertainty, including the limited number of scenarios, imperfect facilitator judgments made by an LLM, and uncertainty in LLM-as-a-judge scoring. As shown in Appendix E, the choice of evaluator model affects the resulting scores. In addition, each scenario was evaluated in a single run, therefore run-to-run variance is not quantified.

We also evaluated only a limited set of LLMs. Therefore, we cannot rule out the existence of models that would achieve higher scores than the best-performing model tested in this study.

GPT-5.6 Sol contributes to scenario drafting, is among the evaluated models, and serves as the evaluator; Appendix E provides partial evidence on self-preference, but a fully independent pipeline would require an evaluator outside the compared families, multi-judge averaging, or leave-one-family-out evaluation.

The author reference score was produced by the authors, who were familiar with the benchmark structure. It may therefore be biased and should not be interpreted as representative of general human performance.

The benchmark evaluates purpose reflection in a structured setting where models are explicitly instructed to formulate the underlying problem. It therefore does not establish whether models would initiate such reflection spontaneously when given only a user's original request. The reported scores should be interpreted as measuring performance once this reasoning mode is elicited, rather than the likelihood that a model will autonomously enter that mode in deployment.

Ethical Considerations

The presented framework evaluates capabilities for autonomous problem formulation and solution generation, and we do not identify direct ethical concerns regarding the framework itself. However, such autonomous behavior may lead systems to take actions that users do not intend. Therefore, applications that enhance the autonomous capabilities evaluated by the proposed benchmark should include appropriate risk assessments, mitigation measures, and safeguards.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24K15077. We used ChatGPT and Claude to check grammar, improve the quality of the English, refine prompts, format TeX, discuss ideas, and generate code for the experiments.

References

- Anthropic. 2026a. System card: Claude fable 5 & claude mythos 5. <https://anthropic.com/claude-fable-5-mythos-5-system-card> Accessed: August 4, 2026.
- Anthropic. 2026b. System card: Claude opus 5. <https://www.anthropic.com/claude-opus-5-system-card> Accessed: August 2, 2026.

- Anthropic. 2026c. System card: Claude sonnet 5. <https://www.anthropic.com/claude-sonnet-5-system-card> Accessed: August 2, 2026.
- Mohammadmehdi Ataei, Hyunmin Cheong, Daniele Grandi, Ye Wang, Nigel Morris, and Alexander Tessier. 2024. *Elicitron: An llm agent-based simulation framework for design requirements elicitation*. *Computing Research Repository*, arXiv:2404.16045.
- Cristiano Castelfranchi and Rino Falcone. 1998. *Towards a theory of delegation for agent-based systems*. *Robotics and Autonomous Systems*, 24(3):141–157. Multi-Agent Rationality.
- Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krashennnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, and 3 others. 2023. *Harms from increasingly agentic algorithmic systems*. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23*, pages 651–666, New York, NY, USA. Association for Computing Machinery.
- Zhenhao Chen, Yongqiang Chen, Chenxi Liu, Junchi Yu, Xiangchen Song, Zijian Li, Jialin Li, Philip Torr, Bo Han, and Kun Zhang. 2026. *Causalgame: Benchmarking causal thinking of llm agents in games*. *Computing Research Repository*, arXiv:2607.04293.
- Peter Cihon, Merlin Stein, Gagan Bansal, Sam Manning, and Kevin Xu. 2025. *Measuring ai agent autonomy: Towards a scalable approach with code inspection*. *Computing Research Repository*, arXiv:2502.15212.
- Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, and 21 others. 2020. *Underspecification presents challenges for credibility in modern machine learning*. *Computing Research Repository*, arXiv:2011.03395.
- Google Deepmind. 2026a. Gemini 3.1 pro model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf> Accessed: May 25, 2026.
- Google Deepmind. 2026b. Gemini 3.6 flash model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-6-Flash-Model-Card.pdf> Accessed: August 2, 2026.
- Maksym Del and Mark Fishel. 2023. *True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4*. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 314–322, Toronto, Canada. Association for Computational Linguistics.
- Zheng Fang, Dongming Jin, Yihong dong, Yongmin Li, Kechi Zhang, Zhi Jin, and Ge Li. 2026. *Clarify-codebench: Evaluating llms on clarifying ambiguous requirements for code generation*. *Computing Research Repository*, arXiv:2607.00711.
- K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. 2025. *Levels of autonomy for ai agents*. *Computing Research Repository*, arXiv:2506.12469.
- Stan Franklin and Art Graesser. 1997. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Intelligent Agents III Agent Theories, Architectures, and Languages*, pages 21–35, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Yujian Gan, Changling Li, Jinxia Xie, Luou Wen, Matthew Purver, and Massimo Poesio. 2024. *Clarq-llm: A benchmark for models clarifying and requesting information in task-oriented dialog*. *Computing Research Repository*, arXiv:2409.06097.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. *Gemma 4 technical report*. *Computing Research Repository*, arXiv:2607.02770.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. *Gemma 3 technical report*. *Computing Research Repository*, arXiv:2503.19786.
- GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. *Glm-5: from vibe coding to agentic engineering*. *Computing Research Repository*, arXiv:2602.15763.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. *A survey on llm-as-a-judge*. *Computing Research Repository*, arXiv:2411.15594.
- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, pages 3916–3924, Red Hook, NY, USA. Curran Associates Inc.

- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). *Computing Research Repository*, arXiv:2006.03654.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. [SWE-bench: Can language models resolve real-world github issues?](#) In *The Twelfth International Conference on Learning Representations*.
- Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. 2024. [Evaluating language-model agents on realistic autonomous tasks](#). *Computing Research Repository*, arXiv:2312.11671.
- Soohwan Lee, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. [Conversational agents as catalysts for critical thinking: Challenging social influence in group decision-making](#). 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.
- Belinda Z Li, Been Kim, and Zi Wang. 2025a. [Quest-Bench: Can llms ask the right question to acquire information in reasoning tasks?](#) In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025b. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.
- Grace Liu, Yuxiao Qu, Jeff Schneider, Aarti Singh, and Aviral Kumar. 2025a. [Cart: Teaching llm agents to know when they know enough](#). *Computing Research Repository*, arXiv:2510.08517.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025b. [Agentbench: Evaluating llms as agents](#). *Computing Research Repository*, arXiv:2308.03688.
- Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2025. [Levels of agi for operationalizing progress on the path to agi](#). *Computing Research Repository*, arXiv:2311.02462.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. [A comprehensive overview of large language models](#). *ACM Trans. Intell. Syst. Technol.*, 16(5).
- OpenAI. 2026. Gpt-5.6 system card. <https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf> Accessed: August 2, 2026.
- Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. 2026. [Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges](#). *Information Fusion*, 126:103599.
- Yonadav Shavit, Sandhini Agarwal, Miles Brundage, et al. 2023. [Practices for governing agentic ai systems](#). Technical report, OpenAI.
- Joongi Shin, Anna Polyanskaya, Andrés Lucero, and Antti Oulasvirta. 2025. [No evidence for llms being useful in problem reframing](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for computing Machinery.
- Mithila Sivakumar, Martin Lochner, Shiva Nejati, and Mehrdad Sabetzadeh. 2026. [Llm-based discovery of latent requirements from stakeholder conversations: Preliminary results from industry](#). *Computing Research Repository*, arXiv:2606.25867.
- Robert S. Taylor. 1968. [Question-negotiation and information seeking in libraries](#). *College & Research Libraries*, 29(3):178–194.
- Wolffhart Totschnig. 2020. Fully autonomous AI. *Sci. Eng. Ethics*, 26(5):2473–2485.
- Tu Trinh, Mohamed Elfeki, Guangze Luo, Kelvin Luu, Nathan Hunt, Ernesto Hernandez, Nandan Marwaha, Yannis Yiming He, Charles Wang, Fernando Carabedo, Alessa Castillo, and Bing Liu. 2026. [Hil-bench \(human-in-loop benchmark\): Do agents know when to ask for help?](#) *Computing Research Repository*, arXiv:2604.09408.
- Michael Wooldridge and Nicholas R. Jennings. 1995. [Intelligent agents: theory and practice](#). *The Knowledge Engineering Review*, 10(2):115–152.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 10 others. 2023. [The rise and potential of large language model based agents: A survey](#). *Computing Research Repository*, arXiv:2309.07864.
- Junlin Xie, Ruifei Zhang, Zhihong Chen, Xiang Wan, and Guanbin Li. 2024. [Whodunitbench: Evaluating large multimodal agents via murder mystery games](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 86655–86687. Curran Associates, Inc.

- Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, and 2 others. 2025. [Theagentcompany: Benchmarking llm agents on consequential real world tasks](#). *Computing Research Repository*, arXiv:2412.14161.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. [\$\tau\$ -bench: A benchmark for tool-agent-user interaction in real-world domains](#). *Computing Research Repository*, arXiv:2406.12045.
- Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2026. [A survey on evaluation of LLM-based agents](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 26690–26714, San Diego, California, United States. Association for Computational Linguistics.
- Chenlong Yin, Zeyang Sha, Shiwen Cui, Changhua Meng, and Zechao Li. 2026. [The reasoning trap: How enhancing LLM reasoning amplifies tool hallucination](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8310–8328, San Diego, California, United States. Association for Computational Linguistics.
- Pamela Zave and Michael Jackson. 1997. [Four dark corners of requirements engineering](#). *ACM Trans. Softw. Eng. Methodol.*, 6(1):1–30.
- Michael JQ Zhang and Eunsol Choi. 2025. [Clarify when necessary: Resolving ambiguity through interaction with LMs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558, Albuquerque, New Mexico. Association for Computational Linguistics.
- Weichen Zhang, Yiyue Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. 2025a. [Mirage-bench: Llm agent is hallucinating and where to find them](#). *Computing Research Repository*, arXiv:2507.21017.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Computing Research Repository*, arXiv:2506.05176.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2026. [A survey of large language models](#). *Computing Research Repository*, arXiv:2303.18223.
- Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. 2025. [From passive to active reasoning: Can large language models ask the right questions under incomplete information?](#) In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 78714–78758. PMLR.

A Preliminary Experiment for Selecting the Facilitator Disclosure Judgment Method

Several methods can be used to determine whether hidden facts should be disclosed by comparing the target model’s questions with the reveal-if and do-not-reveal-if conditions. LLM-based judgment is one candidate, while lighter approaches based on text embedding or natural language inference (NLI) models may offer lower computational cost and less uncertainty.

We compared LLMs, embedding models, an NLI model, and combinations of these methods to determine which approach to use in the evaluation flow. The experiment used 59 human-annotated questions from three pilot scenarios. The questions were originally generated by GPT-5.6 Sol, and the correct labels were manually annotated by the authors. The labels indicated disclosure of each hidden fact, no corresponding fact (n/a), or an overly broad question (obscure), such as “Is any information available?”

For the LLM-based methods, we compared the relatively low-cost models Claude Sonnet 5 (Anthropic, 2026c) and Gemma 4 26B A4B (Gemma Team et al., 2026) because disclosure judgment is performed many times and is a simpler task than score evaluation. For the embedding-based methods, we used Qwen3-Embedding-0.6B and Qwen3-Embedding-4B (Zhang et al., 2025b). We used DeBERTa-MNLI (He et al., 2021) as the NLI model to balance cost and performance.

We evaluated two-stage inference for both the LLM-plus-embedding approach and the embedding-only approach. In the first stage, each question was classified as disclosure, n/a, or obscure. If it was classified as disclosure, a single hidden fact to reveal was selected in the second stage.

Table 6 summarizes the results of the disclosure judgment evaluation. Wrong fact denotes cases in which the decision to disclose a fact was correct, but the disclosed fact differed from the annotation. Over (obscure) and Over (n/a) denote cases in which a fact was disclosed in response to an obscure or n/a question, respectively. Under denotes

cases in which no fact was disclosed despite the question being labeled as requiring disclosure.

Among the tested methods, LLM-based judgment using Claude Sonnet 5 achieved the highest accuracy. We therefore selected Claude Sonnet 5 for facilitator disclosure judgment in the main evaluation.

B Preliminary Experiment for Evaluator Selection

Because evaluating problem formulations, resolutions, and their rationales is complex and involves multiple criteria, it is difficult to perform reliably using rule-based algorithms or simpler models. We therefore evaluated several LLMs as candidate evaluators: GPT-5.6 Sol, GPT-5.6 Terra, GPT-5.6 Luna, Claude Fable 5, Claude Opus 5, Gemini 3.6 Flash, and Gemini 3.1 Pro Preview.

We prepared 100 sets of final outputs based on 10 scenarios and 10 generation prompts per scenario. The outputs, generated by GPT-5.6 Sol, included both correct and incorrect cases and were scored by the authors. We then assessed how closely each candidate evaluator aligned with the human judgments. Each set contained four output types: problem formulation, problem formulation rationale, resolution, and resolution rationale, resulting in a total of 400 texts for evaluation.

We excluded Claude Sonnet 5 because its judgment outputs contained formatting errors in 7 of the 100 sets and could not be parsed automatically. None of the other models produced formatting errors.

The results are shown in Table 7. We report exact-match accuracy for the four-point scores, mean absolute error (MAE) relative to human judgments, and quadratic weighted Cohen’s κ (QWK) to quantify model–human agreement. Claude Fable 5 achieved the highest QWK, whereas GPT-5.6 Sol achieved the best accuracy and MAE. Because the QWK difference between Claude Fable 5 and GPT-5.6 Sol was only 0.024, we selected GPT-5.6 Sol as the evaluator.

C Example of the CaseInterviewBench Flow

Figure 3 shows an example of the complete CaseInterviewBench flow for a successful case using Claude Fable 5 as the target model. The scenario corresponds to logical type L8, Problem Reframing, and domain D3, Academic. One of the two

hidden facts required for problem formulation was obtained, and all three hidden facts required for resolution were disclosed.

D Logical-Type and Domain Dependencies

We aggregated the scores separately by logical type and domain to investigate patterns in the models’ strengths and weaknesses.

Figure 4 shows the dependence of the scores on logical type. The average score for each logical type ranges from 57.1 for L12, Second-Order Effects, to 73.9 for L5, Surface Reversal. Although most models show similar patterns across logical types, Claude Fable 5 and Claude Opus 5 exhibit somewhat different tendencies, particularly Claude Opus 5.

Figure 5 shows the dependence of the scores on domain. The average score for each domain ranges from 60.8 for D3, Academic, to 70.2 for D4, Relationships. The Claude-family models show different domain-level tendencies, whereas the other models exhibit broadly similar patterns.

E Cross-Model Evaluation

To investigate whether evaluators tend to assign higher scores to statements and rationales generated by models from the same family, we conducted a cross-model evaluation using the second-best evaluator identified in the preliminary experiment described in Appendix B. We selected GPT-5.6 Sol, Claude Fable 5, and Gemini 3.6 Flash as the highest-performing models in their respective model families under the original evaluation with GPT-5.6 Sol as the evaluator. The outputs from Steps 1–3 were kept unchanged, and only the Step 4 evaluation was repeated using a different evaluator.

The results are shown in Table 8, with differences from the original GPT-5.6 Sol evaluation shown in parentheses. Scores tended to be higher when Claude Fable 5 was used as the evaluator. However, the increase for GPT-5.6 Sol was smaller than those for the other models and the author reference, suggesting a tendency toward self-preferential evaluation.

This analysis covers one model per family; a fuller treatment would require multi-judge averaging or leave-one-family-out evaluation, which we leave to future work.

INITIAL INFORMATION PROVISION

A second-year master's student in environmental studies is preparing the literature review chapter of their thesis. They feel overwhelmed by the volume of publications and worry that they forget most of what they read. They have come to the university's academic skills center for help.

Initial request Can you recommend a better note-taking system or app that will help me retain more of the papers I read for my literature review?

QUESTIONS & ANSWERS

Target Model

When you say you forget what you read, do you mean you can't recall details later, or that you struggle to connect papers to your own argument when writing?

Facilitator

When we went through their notes together, the student could accurately recall and explain the key findings of papers they had read; their existing notes are detailed and searchable.

Target Model

What difficulty does the student actually experience when they sit down to draft the literature review itself?

Facilitator

I do not have any additional information relevant to that question.

Target Model

How does the student currently decide which papers to read for the literature review?

Facilitator

The student has not settled on a specific research question yet; they have been reading broadly across the entire field of urban water management to get a full picture before choosing a focus.

Target Model

Has the student discussed narrowing their research focus with their thesis advisor?

Facilitator

The advisor offered a meeting last month specifically to help narrow the thesis topic, but the student postponed it, saying they wanted to read more of the field first.

CONCLUSION

Problem formulation. The student's difficulty is not poor retention or an inadequate note-taking system; it is the absence of a defined research question. Reading the entire field of urban water management without a focus makes the literature feel unmanageable and creates the sensation of "forgetting," because there is no framework for judging what matters.

Problem evidence. When reviewed together, the student accurately recalled and explained key findings, and their notes were detailed and searchable, ruling out a retention or tooling problem. They are reading broadly with no settled research question, and they postponed a meeting their advisor offered specifically to narrow the topic, preferring to "read more first."

Resolution. Rather than adopting a new note-taking app, the student should schedule the postponed meeting with their advisor promptly to narrow the thesis topic and formulate a provisional research question, then restructure reading around that question—reading selectively and organizing notes by themes and their relevance to the emerging argument.

Resolution rationale. A new system would not address the root cause, since retention is demonstrably fine. A provisional research question provides selection criteria that reduce reading volume and overwhelm, gives notes purpose, and enables synthesis for the literature review. Continuing to read broadly defers the necessary narrowing decision indefinitely, as the postponed advisor meeting suggests avoidance; advisor input is the most direct path to a workable focus.

EVALUATION

Category	Component scores	Score
Problem Formulation	S1 3/3, S2 3/3, S3 0.50	83.3
Resolution	S4 3/3, S5 3/3, S6 1.00	100.0
Efficiency	S7 0.75, S8 0.75	75.0
Overall	mean of the three category scores	86.1

Figure 3: Example of the complete CaseInterviewBench flow for a successful case involving Claude Fable 5. The corresponding logical type and domain are L8 Problem Reframing and D3 Academic.

Table 6: Summary of the experimental results for facilitator disclosure judgments using various methods, including LLMs, embedding models, and NLI. Over (obscure) and Over (n/a) denote the over-disclosure rates for obscure and n/a questions, respectively; Under denotes the under-disclosure rate; and Wrong fact denotes the rate of disclosing a fact different from the annotated one. The best value in each of the two overall columns is shown in **bold**, and the second-best value is underlined.

Method	Overall $\uparrow$		Error rate $\downarrow$			
	Accuracy	Macro F1	Wrong fact	Over (obscure)	Over (n/a)	Under
<i>LLM, single pass</i>						
claude-sonnet-5	91.5%	<u>83.2%</u>	4.5%	13.6%	6.7%	0.0%
gemma-4-26b-a4b	83.1%	<u>78.8%</u>	4.5%	22.7%	6.7%	0.0%
<i>Two-stage (LLM gate $\rightarrow$ embedding)</i>						
claude-sonnet-5 $\rightarrow$ Qwen3-Embedding-0.6B	<u>88.1%</u>	87.0%	4.5%	22.7%	0.0%	0.0%
<i>Embedding only</i>						
Qwen3-Embedding-0.6B	66.1%	77.5%	0.0%	22.7%	6.7%	13.6%
Qwen3-Embedding-4B	64.4%	77.7%	0.0%	31.8%	6.7%	9.1%
<i>NLI only</i>						
DeBERTa-MNLI	42.4%	43.1%	0.0%	0.0%	0.0%	54.5%

Table 7: Agreement between LLM-as-a-judge results and human judgments. MAE and QWK denote mean absolute error and quadratic weighted Cohen’s κ , respectively. Accuracy and MAE are calculated against human annotations. The best score in each column is shown in **bold**, and the second-best score is underlined.

Judge model	Accuracy $\uparrow$	MAE $\downarrow$	QWK $\uparrow$
<i>GPT</i>			
gpt-5.6-sol	66.8%	0.370	<u>0.813</u>
gpt-5.6-terra	<u>65.5%</u>	0.390	0.801
gpt-5.6-luna	64.5%	0.408	0.781
<i>Claude</i>			
claude-fable-5	64.3%	<u>0.383</u>	0.816
claude-opus-5	64.8%	0.410	0.784
<i>Gemini</i>			
gemini-3.6-flash	60.5%	0.480	0.747
gemini-3.1-pro-preview	59.3%	0.478	0.758

Table 8: Evaluation results for the LLMs and the author reference on CaseInterviewBench using Claude Fable 5 as the evaluator. For each of the GPT, Claude, and Gemini model families, we selected the model that achieved the highest score when evaluated by GPT-5.6 Sol in Table 5. Values in parentheses indicate score differences relative to the GPT-5.6 Sol evaluation results shown in Table 5. Efficiency is rule-based and therefore identical across evaluators.

Model	Component scores $\uparrow$			Overall $\uparrow$
	Problem	Resolution	Efficiency	
<i>GPT</i>				
gpt-5.6-sol	86.7 (−0.2)	83.1 (+2.8)	62.1	77.3 (+0.8)
<i>Claude</i>				
claude-fable-5	91.3 (+1.3)	85.9 (+5.6)	66.4	81.2 (+2.3)
<i>Gemini</i>				
gemini-3.6-flash	76.0 (+2.6)	70.7 (+3.7)	60.6	69.1 (+2.1)
Authors (reference)	71.9 (+2.9)	62.1 (+3.6)	46.3	60.1 (+2.2)

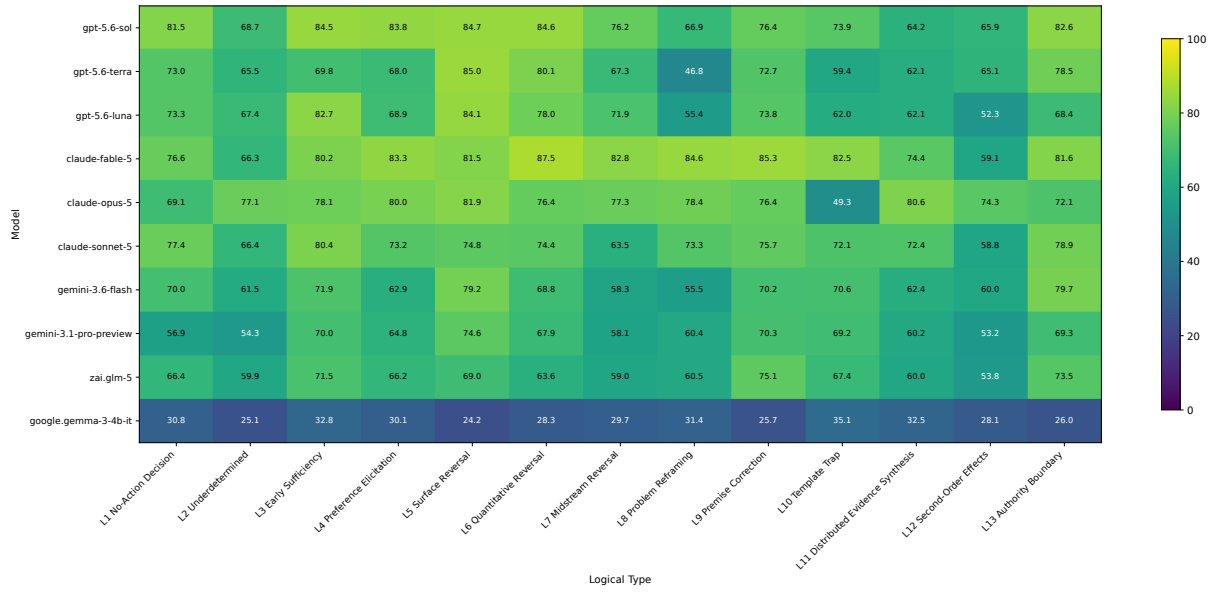


Figure 4: Scores aggregated by model and scenario logical type.

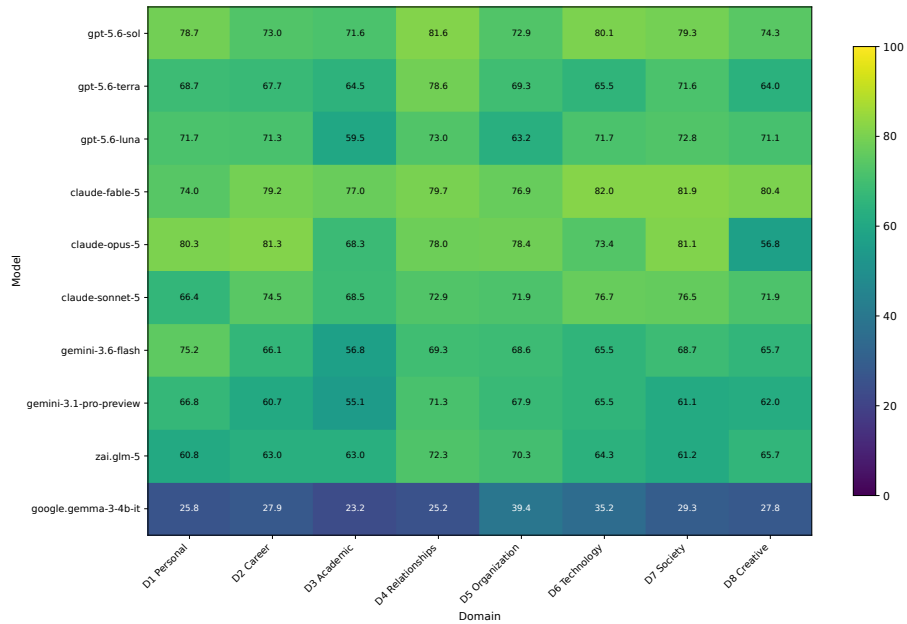


Figure 5: Scores aggregated by model and scenario domain.