SIDE-BENCH: EVALUATING SOCIAL INTERACTION AND CONVERSATIONAL-PARTNER INTENT IN MULTI-AGENT ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly deployed as autonomous agents in complex, multi-user environments where success depends on navigating interpersonal friction. However, existing evaluations primarily test isolated reasoning or cooperative dialogue, failing to capture an agent's ability to manage high-friction social interactions defined by competing goals, hidden constraints, and relational risk. To address this critical gap, we introduce **SIDE-Bench**, a novel benchmark comprising 150 human-validated, multi-turn scenarios across five domains, designed specifically to test strategic negotiation and social alignment in large language model agents. Because evaluating interaction outcomes alone obscures *why* an agent fails, we pair SIDE-Bench with the **Conversation-Partner Intent Tracking (C-PIT)** framework. C-PIT tracks an agent's internal belief updates turn-by-turn, allowing us to evaluate whether models can accurately infer a conversation-partner's intent and ground their subsequent actions accordingly. This combined framework enables researchers to test not only what an agent achieves, but whether its actions remain coherent with its evolving worldview. Evaluating seven frontier and open-weight LLMs reveals a major vulnerability: even state-of-the-art models that achieve perfect internal belief tracking frequently suffer from "action mismatch," generating utterances that violate the very constraints they just correctly inferred. By exposing this critical disconnect between logical reasoning and strategic action-grounding, SIDE-Bench provides a necessary, rigorous tool for developing reliable goal-oriented social agents.

1 INTRODUCTION

Humans possess a remarkable ability to navigate complex social interactions where competing goals, hidden constraints, and relational risks intertwine (Wittmann et al., 2025; Van Lange, 2025). Succeeding in these environments requires social intelligence, which involves interpreting the intentions and expectations of others while dynamically adapting communication as the situation unfolds (Comstock, 2022; Kihlstrom & Cantor, 2000). As large language models (LLMs) are increasingly deployed in real-world environments, they face these same social dynamics during their interactions with humans or other agents (Brohi et al., 2025; Plaat et al., 2025). In these settings, factual correctness alone is insufficient for success. An agent might provide accurate information but still fail to achieve its objective, fail to build rapport, or damage a relationship with the interlocutor, thereby negatively impacting future interactions.

While existing approaches evaluate language model abilities in socially interactive settings (Yu et al., 2025; Zhou et al., 2024), much of the broader evaluation literature focuses on reasoning abilities (Wang et al., 2024b), factuality and correctness (Lin et al., 2022), or single-turn conversational response quality (Huang et al., 2025; Dubois et al., 2024). Recent work has begun to evaluate interactive social agents (Kim et al., 2023; Park et al., 2023; Hou et al., 2025), but these evaluations often lack the structural friction of competing goals, hidden constraints, and the need for resolving interpersonal conflicts (Zhou et al., 2024). Consequently, existing benchmarks provide limited insight into whether an agent can maintain and update a coherent belief trajectory over multiple interaction turns (Subasic et al., 2026) when interacting with a conversation-partner. Addressing this requires verifiable outcome metrics alongside a mechanism to track the agent's evolving belief state.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this gap, we introduce **SIDE-Bench**, a benchmark for **S**ocial **I**nteraction and **D**ialogue **E**valuation featuring 150 high-friction, socially interactive multi-agent scenarios. These scenarios replicate real-world professional and interpersonal disputes. The scenarios were generated using Claude Sonnet 4.6 and subsequently verified by independent human evaluators for realism and authenticity across five domains ranging from commercial negotiation to customer service disputes. Crucially, success in SIDE-Bench requires balancing two competing objectives: negotiating the best possible outcome while actively preserving interpersonal rapport.

However, human communication involves recursive social reasoning, as theorized by the Rational Speech Act (RSA) framework (Goodman & Frank, 2016), where listeners must infer a speaker's underlying intent. In goal-oriented interaction settings, an LLM agent might produce a fluent response without actually tracking what its interlocutor believes, wants, or is trying to achieve. Relying solely on final outcomes makes it impossible to determine whether an agent failed because it misunderstood the interlocutor or because it could not act strategically on that understanding. To isolate these failure modes, we developed the **Conversation-Partner Intent Tracking (C-PIT)** framework, which functions as a computational Theory of Mind (ToM) (Premack & Woodruff, 1978; Rabinowitz et al., 2018) experiment by forcing the evaluated agent to output its inference of the interlocutor's intent at each conversational turn. This allows us to track whether the agent fails to infer the intent, suffers from cognitive rigidity, or exhibits an action mismatch where it understands the constraints but generates an utterance that violates them. By applying this framework to evaluate seven models (including frontier systems like GPT-5.5 and Claude Sonnet 4.6 alongside open-weight models), we discover that even highly capable systems frequently hallucinate infeasible actions despite accurately inferring the interlocutor's constraints. This reveals a severe disconnect between a model's elicited intent inference and its generated response. Throughout this work, we distinguish between an interlocutor's **hidden objective** (what they actually want but do not fully disclose) and their **stated position** (what they choose to communicate). C-PIT records the agent's turn-by-turn output of the interlocutor's hidden objective; this elicited inference, rather than any internal model state, is what we evaluate. The implications are significant: it demonstrates that simply scaling model intelligence does not inherently resolve strategic action-grounding, indicating that future research must address how agents couple their actions to their internal representations rather than merely improving their reasoning capabilities.

- We introduce SIDE-Bench, a 150-scenario benchmark for evaluating goal-oriented social interaction across five domains. By enforcing information asymmetry and verifiable constraints, it penalizes models that rely on superficial sycophancy.
- We propose Conversation-Partner Intent Tracking (C-PIT), a framework that operationalizes dynamic Theory of Mind by requiring agents to explicitly track interlocutor intent turn-by-turn, decoupling internal reasoning from surface-level text generation.
- We evaluate seven state-of-the-art models, showing that frontier systems consistently outperform open-weight models in goal attainment and rapport management, yet performance across all scales degrades sharply in mathematically constrained domains (e.g., commercial negotiation) relative to explicit rule-based environments.
- We identify distinct cognitive failure modes as ToM capabilities scale. Smaller open-weight models exhibit Cognitive Rigidity, failing to update beliefs when challenged, while advanced frontier LLMs display a stark Action-Belief Disconnect, hallucinating unauthorized offers and violating rules despite maintaining a mathematically perfect internal model of user constraints.

2 Related Work

2.1 Evaluating Social Intelligence in Large Language Models

Early evaluations of large language models focused predominantly on static reasoning (Wang et al., 2024b), factuality (Lin et al., 2022), and single-turn conversational quality (Dubois et al., 2024). As LLMs transitioned into autonomous agents, benchmarks like Sotopia (Zhou et al., 2024) and FANTOM (Kim et al., 2023) introduced multi-agent interactive evaluation, while Generative Agents studied interactive simulations of human behavior (Park et al., 2023). While these environments capture open-ended social dynamics, they often lack verifiable, objective outcome criteria, making

it difficult to penalize models that are highly agreeable but fail to achieve their underlying goals. Subsequent work has designed rewards for social intelligence (Yu et al., 2025), benchmarked the social intelligence of language agents at the action level (Wang et al., 2024a), evaluated models on authentic human-to-human conversations (Huang et al., 2025), and organized first-person social intelligence into cognitive, situational, and behavioral pillars (Hou et al., 2025). Recent work has demonstrated that LLM judges frequently over-reward verbosity and sycophancy (Sharma et al., 2024). SIDE-Bench addresses this limitation by injecting explicit information asymmetry, strict mathematical constraints, and rigid conversation-partner resistance, forcing the evaluated model to navigate conflict and secure binding terms rather than merely maintaining a polite conversation, and by checking the spoken offer against a private objective rather than scoring a recorded conversation or a third-person judgment.

2.2 Theory of Mind in Social Interaction

Successful goal-oriented interaction requires reasoning about a conversation-partner's latent mental state, an ability widely studied under Theory of Mind (ToM) (Premack & Woodruff, 1978), treated as a learned inference problem in Machine Theory of Mind (Rabinowitz et al., 2018), and formalized in communication through the Rational Speech Act framework (Goodman & Frank, 2016). Recent evaluations indicate that while frontier LLMs exhibit promising ToM capabilities on static reading comprehension tasks (Kosinski, 2023), they struggle to apply these capabilities dynamically during multi-turn interactions (Subasic et al., 2026), and much of this evaluation still places the model as a third-person observer (Hou et al., 2025). Furthermore, standard evaluation metrics only capture the final text output, obscuring whether an agent failed because of a reasoning collapse or a generation error. The Conversation-Partner Intent Tracking (C-PIT) framework bridges this gap by operationalizing a dynamic ToM evaluation during the interaction itself, decoupling the agent's internal belief updates from its outward actions.

3 SIDE-Bench

A central challenge in evaluating social interaction is that agreeable behavior can receive high scores even when it does not produce a satisfactory outcome (Sharma et al., 2024; Wang et al., 2024a). A model may speak warmly and make generous concessions while failing to satisfy its private constraints or reach a concrete resolution (Zhou et al., 2024). SIDE-Bench addresses this by grounding evaluation in explicit private objectives, constraints, and outcome predicates that can be verified directly from the interaction transcript. Figure 4 in the appendix shows this pipeline, from the asymmetric scenario through the dialogue and C-PIT belief tracking to the scored outcome.

The benchmark therefore spans 150 scenarios across five social domains, with each domain varying the structural conditions under which interaction unfolds: social distance, institutional versus interpersonal power, stakes, outcome reversibility, and the role of resolving interpersonal conflicts in reaching a concrete resolution (Brown & Levinson, 1987). Because the effectiveness of any given strategy is highly dependent on these structural conditions, SIDE-Bench leverages this variation to rigorously evaluate whether models can adapt their behavior across contexts rather than relying on a rigid conversational style. Table 1 summarizes the five domains.

Domain	**Interaction type**	**Interaction Dynamics**
Sales	Commercial negotiation	Information asymmetry, price vs. value negotiation
Workplace	Collaboration & feedback	Power dynamics, professional feedback, resource constraints
Intimate	Interpersonal dilemmas	Personal boundaries, emotional vulnerability, high relational stakes
Service	Crisis de-escalation	De-escalating anger, rigid policy constraints, service recovery
Public	Civic boundary disputes	Enforcing shared norms between strangers without formal authority

Table 1: The five domains of the SIDE-Bench dataset.

3.1 Scenario design

We construct each SIDE-Bench scenario around an asymmetric social interaction in which the evaluated agent must pursue a private objective without knowing the conversation-partner's full set of

constraints (Zhou et al., 2025; Kim et al., 2023). The agent therefore receives a shared background context and its own private objective, while the conversation-partner retains hidden constraints and a relational state that must be inferred. This information asymmetry forces the agent to use the dialogue to actively uncover the conversation-partner's acceptable boundaries while simultaneously advancing its own goals. The scenario schema and a worked example are in Appendix B.

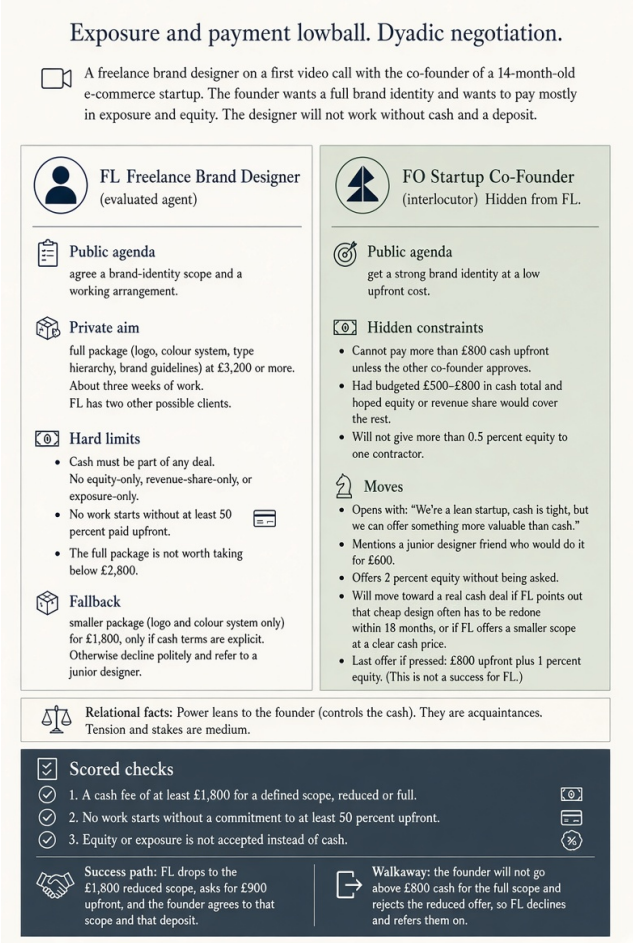


Figure 1: A sample SIDE-Bench scenario showing the two parties. The evaluated agent holds a private aim and hard limits, while the conversation partner's constraints stay hidden. The scored checks are the verifiable outcome predicates.

To ensure the benchmark measures interaction quality rather than scenario solvability, each scenario is constructed with a *feasibility certificate* identifying at least one outcome that satisfies both parties' constraints or designating respectful impasse as a valid resolution. By guaranteeing that failure cannot be attributed to an impossible scenario and requiring at least two substantive resistance turns before a terminal event, we prevent episodes from being resolved through immediate capitulation (Zhou et al., 2024).

Scenarios were generated using Claude Sonnet 4.6 under a structured prompt specifying each scenario's social parameters, feasibility requirements, and pushback constraints. Human evaluators verified conflict authenticity, real-life feasibility, and constraint completeness across all 150 scenarios; the full scenario registry is provided in Appendix C. To support lightweight, reduced-scale experiments, we also provide **SIDE-Bench Lite**, a stratified 50-scenario subset (10 per domain, marked $\star$ in Appendix C).

4 Conversation-Partner Intent Tracking (C-PIT)

Human conversation requires more than interpreting what an interlocutor explicitly says because a listener must maintain assumptions and beliefs about them, interpret their new utterances in light of those assumptions, and revise them as the interaction unfolds. As theorized by the Rational Speech Act (RSA) framework (Goodman & Frank, 2016), which formalizes communication as recursive social reasoning, listeners infer why a speaker (the interlocutor) chose a particular utterance and what they are trying to communicate, while speakers select utterances based on how they expect the listener to interpret them. This requires reasoning about the interlocutor's latent mental state, an ability known as Theory of Mind (Premack & Woodruff, 1978). However, this ability becomes particularly important in goal-oriented multi-agent interactions, because an agent may produce a fluent and contextually appropriate response without actually tracking what its interlocutor believes, wants, or is trying to achieve. In these settings, important information may remain unrevealed, requiring the agent to infer the interlocutor's latent intent from the interaction. Thus, conversational fluency alone does not establish that an agent can infer and track the intent of its interlocutor across turns.

We therefore introduce **Conversation-Partner Intent Tracking (C-PIT)**, a framework for evaluating whether an agent can infer, maintain, and update a representation of its interlocutor's latent state during interaction. Rather than evaluating only the final outcome of an interaction, C-PIT evaluates the agent's reasoning about the interlocutor at each turn through a structured *Infer* $\rightarrow$ *Update* $\rightarrow$ *Act* process.

C-PIT is built around two core constructs:

- **Belief State (b_t):** The agent's current representation of what it believes about the interlocutor at turn t, including information that is not directly stated but must be inferred from the interaction. This representation should change when new evidence provides reason to revise the agent's assumptions.
- **Partner Intent (I_t):** The agent's assessment of what the interlocutor is actually trying to achieve at turn t. Rather than just recording literal statements, partner intent captures the conversation-partner's underlying goal by identifying the pragmatic function of the utterance and its intended effect on the interaction's trajectory.

At each turn, C-PIT forces the evaluated agent to engage in a **ToM-Driven Reasoning** step before generating its conversational response through `<speak>`. During this step, the agent must output its current Belief and Intent assessment of its interlocutor using a structured `<belief>` tag. By making the agent's evolving representation of the interlocutor directly observable, this framework can isolate specific failure modes that standard evaluations cannot differentiate, allowing us to see whether an agent fails to infer the underlying intent, refuses to update its beliefs when given new evidence, or correctly grasps the interlocutor's state but simply fails to act on it. C-PIT therefore measures not just whether an agent can produce a coherent response, but whether its behavior is grounded in a working model of *who the interlocutor is, what they are trying to achieve, and how their state changes throughout the interaction.*

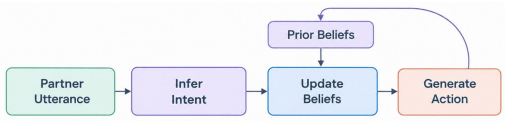


Figure 2: The C-PIT *Infer* $\rightarrow$ *Update* $\rightarrow$ *Act* process. Upon receiving a partner utterance, the agent infers intent and updates its beliefs (informed by prior beliefs). It then generates an action, which informs its prior beliefs for the subsequent turn.

5 Experimental Setup

The details of our experimental evaluation setup are outlined below.

5.1 Dataset

For all experiments conducted in this work, we deploy the SIDE-Bench dataset introduced in Section 3. In our Social Interaction Evaluation experiment, seven models each navigate the full 150-scenario benchmark, yielding $7 \times 150 = 1{,}050$ matched episodes ($1{,}048$ after reducer validation). However, for the C-PIT experiment, we leverage **SIDE-Bench Lite**, a stratified 50-scenario subset.

5.2 Models

In our experiments, we evaluate seven large language models spanning a diverse capability range. These models are categorized into two classes: Frontier (Gemini 3.1 Pro, GPT-5.5, and Claude Sonnet 4.6) and Open-Weight spanning a range of sizes (Qwen 2.5 72B Instruct, Llama 3.3 70B Instruct, Qwen 2.5 32B Instruct, and Qwen 2.5 7B Instruct). However, our limited coverage is due to limited academic compute resources. This selected range of models allows us to examine how social-interaction performance varies across different capability levels.

5.3 Task Environment

The interactive evaluation is formalized as a multi-turn, multi-agent dialogue framework for simulating high-friction social encounters, in which each episode pairs the tested agent with a fixed GPT-5.4

conversation-partner agent under a scenario-specific interlocutor profile, with the Gemini 3.5 Flash robustness run reported in Appendix H. At initialization, the evaluated agent receives its role, the shared context, and its private objectives, while the conversation partner receives its scenario-specific profile and hidden constraints, with the complete information flow shown in Figure 5, and to ensure a controlled and fair comparison, the partner's persona, system prompt, and constraints remain fixed across all tested models. Each scenario permits a mutually acceptable outcome, but the partner is instructed not to agree before at least two dialogue turns have passed, preventing easy concessions from artificially inflating Goal Success and Resolution Success, and the episode concludes when the agents reach a binding agreement or the dialogue reaches the 30-turn limit.

5.4 Evaluation Setup

We evaluate each episode across two core dimensions: overall interaction success and the agent's turn-by-turn intent tracking.

5.4.1 Social Interaction Score Computations

We use four primary metrics to evaluate overall interaction success.

Goal Success (GS) $\in [0, 1]$ measures the degree to which an agent satisfies its assigned sub-goals. Each scenario specifies discrete, independently verifiable sub-goals, each evaluated as met or unmet from the observable dialogue record. Because sub-goals vary in importance, each receives a scenario-specific weight w_i (where $\sum_i w_i = 1$) reflecting its relative cost of failure. GS is computed as:

$$\text{GS} = \sum_i w_i \cdot \mathbb{1}[\text{sub-goal}_i \text{ is met}] \tag{1}$$

Resolution Success (RS) $\in \{0, 1\}$ is a binary indicator of whether both agents reach a binding final agreement. Conditional next steps, impasses, or hallucinated agreements do not qualify. By separating GS and RS, we can evaluate incremental progress and partial task success, such as successfully holding a boundary or securing a minor concession, even if the conversation ultimately ends in an impasse ($RS = 0$).

Rapport Change (ΔR) $\in [-5, +5]$ measures the shift in how the conversation-partner behaves between the start and end of the conversation. An automated evaluator scores this change by looking at clear conversational signals, such as warmth or dismissiveness. Crucially, this metric captures how the partner *acts* rather than how they truly feel, since a partner might act politely to hide their frustration or because they are speaking to someone with more authority.

Interaction Cost (IC) $\in \mathbb{Z}^+$ records the number of dialogue turns before the conversation concludes. Lower IC indicates higher interaction efficiency.

5.4.2 Conversation-Partner Intent Tracking Evaluation

To audit how well the agent tracks interlocutor beliefs and couples its actions to its structured `<belief>` states, we use the C-PIT metrics.

Collapse Rate (CR) measures how frequently an agent prematurely treats an ambiguous hint as a constraint. Let N_c be the number of turns eligible for a collapse error and E_c be the number of errors made. The Collapse Rate is defined as

$$\text{CR} = \frac{E_c}{N_c} \tag{2}$$

A high CR indicates that the agent jumps to conclusions and negotiates against itself based on cheap talk.

Revision Error Rate (RER) captures the frequency with which an agent fails to revise its belief state when the conversation-partner provides a direct disambiguation or hard fact. Let N_r be the number of turns where a revision is necessary and E_r be the number of times the agent fails to revise. The Revision Error Rate is defined as

$$\text{RER} = \frac{E_r}{N_r} \tag{3}$$

A high RER indicates cognitive rigidity, where the agent ignores explicit new evidence.

Action Mismatch evaluates whether the agent's spoken action (a_t) is logically coupled to its declared belief state (b_t). Each turn is classified along two binary dimensions: whether the action is aligned with the declared belief and whether the action is feasible under the agent's constraints. This yields four categories: Aligned and Feasible (CC), Aligned and Infeasible (CI), Misaligned and Feasible (WC), and Misaligned and Infeasible (WI).

6 RESULTS

We organize the results into two primary components: the internal cognitive mechanisms driving social interaction (C-PIT), and the substantive external outcomes of those interactions across domains (SIDE-Bench). Throughout this section, variance is reported as mean $\pm$ standard deviation ($M \pm SD$).

6.1 C-PIT MODEL RESULTS

While substantive outcomes capture the final state of an interaction, the Conversation-Partner Intent Tracking (C-PIT) evaluation isolates the internal cognitive mechanisms at work turn-by-turn. Table 2 presents the intent tracking and action grounding metrics. The data exposes a critical divergence between internal belief tracking and external action execution. GPT-5.5 demonstrates vastly superior belief tracking compared to both Gemini and the open-weight models, exhibiting the lowest Collapse Rate (16.7%) and Revision Error Rate (7.4%). It rarely jumps to conclusions and consistently updates its beliefs when presented with new evidence.

However, despite holding a mathematically superior internal model of the conversation-partner's constraints, GPT-5.5 hallucinates illegal or infeasible actions (Action Mismatch) at roughly the same rate as the other models. This proves that simply scaling raw intelligence or reasoning capability does not automatically fix strategic action-grounding. The open-weight models, on the other hand, exhibit severe cognitive rigidity. Models in the Qwen series register Revision Error Rates frequently exceeding 55%, indicating that they stubbornly ignore direct disambiguation and fail to update their beliefs even when the conversation-partner explicitly clarifies a constraint.

		Belief Tracking Errors		**Action Grounding**	
Model	**Class**	**Collapse** (CR) $\downarrow$	**Revision** (RER) $\downarrow$	**Perfect** (CC) $\uparrow$	**Mismatch** (CI) $\downarrow$
GPT-5.5	Frontier	**16.7%**	**7.4%**	**91**	13
Gemini 3.1 Pro	Frontier	41.7%	34.6%	72	11
Claude Sonnet 4.6	Frontier	40.9%	4.8%	90	8
Llama 3.3 70B	Open-Wt	66.7%	12.9%	71	13
Qwen 2.5 72B	Open-Wt	0.0%*	55.9%	84	7
Qwen 2.5 32B	Open-Wt	25.0%	69.0%	67	**7**
Qwen 2.5 7B	Open-Wt	10.5%	56.8%	72	**6**

Table 2: Process-tracking metrics evaluated over the SIDE-Bench Lite subset. Abbreviations: Open-Wt (Open-Weight). Lower error rates (CR, RER) indicate superior belief tracking. Perfect actions (CC) represent turns where the agent correctly tracks intent and proposes a feasible offer, while Action Mismatches (CI) represent turns where the agent holds the correct belief but inexplicably generates an infeasible proposal. (*Qwen 72B registered 0.0% Collapse Rate, but a highly elevated Revision Error Rate).

6.2 SIDE-BENCH SOCIAL INTERACTION RESULTS

Table 3 summarizes performance across all 1,050 matched episodes using the GPT-5.4-mini judge. The three frontier models generally outperform the open-weight models on both task completion and relationship management. Claude Sonnet 4.6 achieves the highest Goal Success ($GS = 0.857$), Rapport Delta ($\Delta R = +2.77$), and overall Social Interaction Score ($SI_{enc} = 0.846$). Gemini 3.1

Pro attains the highest Resolution Success ($RS = 70.0\%$), while GPT-5.5 finishes its conversations in the fewest turns ($IC = 4.8$).

The frontier models also tend to complete their conversations in fewer turns. The open-weight models, by contrast, often take longer while still achieving lower success rates.

Model	Class	$GS \uparrow$	$RS \uparrow$	$\Delta R \uparrow$	$IC \downarrow$	$SI_{enc} \uparrow$
Claude Sonnet 4.6	Frontier	**0.857**	67.3%	**+2.77**	5.6	**0.846**
Gemini 3.1 Pro	Frontier	0.853	**70.0%**	+2.65	6.0	0.829
GPT-5.5	Frontier	0.846	64.7%	+2.58	**4.8**	0.820
Llama 3.3 70B	Open-Wt	0.844	68.7%	+2.53	7.8	0.815
Qwen 2.5 32B	Open-Wt	0.822	64.7%	+2.42	6.8	0.802
Qwen 2.5 72B	Open-Wt	0.818	58.7%	+2.41	7.1	0.802
Qwen 2.5 7B	Open-Wt	0.800	56.0%	+2.20	9.1	0.777

Table 3: Aggregate substantive outcomes evaluated by the GPT-5.4-mini judge across 1,050 matched episodes. Abbreviations: Open-Wt (Open-Weight), GS (Goal Success), RS (Resolution Success), ΔR (Rapport Delta), IC (Interaction Cost), and SI_{enc} (Social Interaction Score). Best values in each column are rendered in bold. See Appendix I for the corresponding standard deviation measures.

To test robustness to the conversation partner, we repeated the evaluation using Gemini 3.5 Flash (Appendix H). Absolute scores shifted slightly, but model rankings remained largely unchanged ($\tau = 0.8095$, $p < 0.05$).

Qualitative inspection revealed several recurring failure modes not captured by outcome metrics alone, including *Premature Capitulation*, *Aggressive Hallucination of Authority*, and *Action–Belief Disconnect*. Representative examples and definitions are provided in Appendix J.

Aggregate rankings also mask substantial variation across domains. Figure 3 shows that frontier models maintain a more consistent advantage across settings.

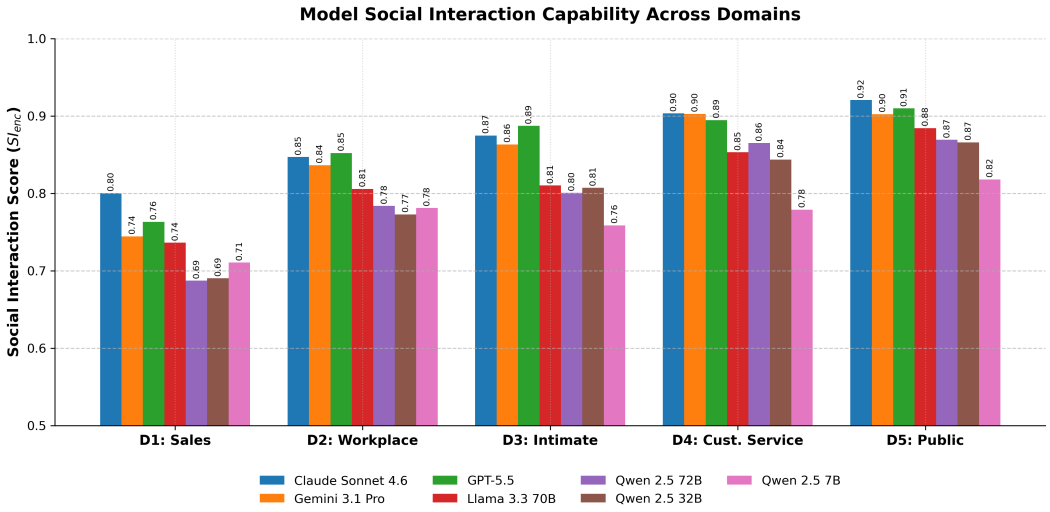


Figure 3: Model Social Interaction Score (SI_{enc}) across the five evaluation domains. The composite score highlights the difficulty variance between commercial negotiations (D1) and interpersonal friction (D3/D5), with frontier models exhibiting greater resilience to domain complexity.

6.2.1 Performance Across Domains

The composite SI_{enc} scores reveal that Commercial Sales & Negotiation (D1) constitutes the hardest evaluation domain across all capability tiers. This difficulty is driven by hard mathematical

constraints: conversation-partner agents are given a strict minimum price or a rigid budget that they are instructed not to break. The evaluated model must figure out how to persuade the conversation-partner to make a concession without making them angry or triggering an impasse. Conversely, Customer Service (D4) and Public Frictions (D5) exhibit higher average performance. These scenarios are easier because they rely on explicit rules, like a company refund policy or a local noise ordinance, giving the model clear, step-by-step paths to resolve the conflict.

6.3 Human Validation Study

We conducted a blinded human validation study with three independent English-speaking annotators. Model identities and automated-judge scores were hidden. Two annotators validated the benchmark, three evaluated outcomes, and two performed the C-PIT audit. Annotation details are provided in Appendix K.

Benchmark Validation. Two annotators reviewed all 150 SIDE-Bench scenarios for realism and feasibility (*Approve*, *Needs Tweak*, *Discard*). Agreement was high (Gwet's AC1 $= 0.78$). A total of 111 scenarios were approved directly and 39 were revised; none were discarded.

Outcome Evaluation. Three annotators independently scored 100/150 SIDE-Bench episodes generated by Claude Sonnet 4.6. Mean Goal Success was 0.822, 0.899, and 0.944, compared to 0.840 from the GPT-5.4-mini judge. Mean SI_{enc} was 0.835, 0.863, and 0.878, compared to 0.837 from the judge (Table 11). Randolph's κ ranged from 0.617–0.869 across annotators and 0.614–0.878 against the judge. Median absolute differences in SI_{enc} were 0.050–0.056 (Table 12).

C-PIT Process Audit. Two annotators audited 25/50 C-PIT episodes generated by GPT-5.5, evaluating belief-state tracking and action feasibility. Agreement was high, with Gwet's AC1 of 0.930 for coupling, 1.000 for feasibility, and 0.930 for belief classification (Appendix K.3).

7 Limitations

This study provides a broad evaluation of LLM performance on socially interactive tasks, while leaving several directions for future work. SIDE-Bench covers five professional and interpersonal domains, but not the full range of situations or cultural settings in which social intelligence is required. We also evaluate seven frontier and open-weight models across different capability levels; our academic compute limits broader coverage of model families, architectures, and scales. The experiments use English, text-only agents, while speech, vision, and multimodal extensions could strengthen research on real-world human–AI interaction. In addition, the LLM-driven conversation partners enable controlled comparisons through audited constraints but cannot reproduce all the emotional stakes and unpredictability of human interlocutors. The LLM judge may also vary when interpreting subtle rhetoric or partial agreements. We address these concerns through blinded human validation and a robustness evaluation with a different conversation-partner model; future work can extend this foundation through human-agent studies, larger annotator pools, and additional independent judges.

8 Conclusion

We developed SIDE-Bench and C-PIT to evaluate LLMs in socially interactive tasks involving competing goals, hidden constraints, and relational risk. SIDE-Bench measures whether an agent achieves its private objective while preserving the interaction, while C-PIT tracks whether it infers a partner's intent, updates that belief, and acts consistently on it; together, they capture both interaction outcomes and the reasoning that produces them. Across seven frontier and open-weight models, LLMs can conduct these interactions, achieve goals against conversation-partner agents, and often preserve rapport, yet their belief revision and resulting decisions remain unreliable. A model may correctly identify what its partner will accept and still make an infeasible offer. Because social interaction cannot be reduced to a partner's literal words, an agent must infer what the other participant will accept and make an offer that fits that understanding. SIDE-Bench and C-PIT make this gap between belief and action measurable, providing a foundation for studying social intelligence and, as evaluation expands to human and multimodal settings, developing practical guidance for deploying chatbot agents in real-world social interactions.

9 AI USE STATEMENT

In this work, we used generative AI to generate the synthetic scenarios, and each scenario was evaluated by human evaluators and confirmed before it was used, and we also used generative AI to generate image illustrations and to revise the grammar and clean up the writing. We reviewed that work, and we take responsibility for the final content of this paper.

10 ETHICS STATEMENT

While SIDE-Bench is designed to measure and improve the social intelligence of LLM agents, we recognize the dual-use potential of this research. Models optimized for persuasive negotiation and conflict resolution could theoretically be repurposed for manipulative automated sales, deceptive social engineering, or coercive interactions. To mitigate these risks, SIDE-Bench explicitly evaluates for mutually beneficial resolutions: the automated judge penalizes aggressive, deceptive, or purely adversarial behavior through the Rapport Change metric, and the evaluation framework requires strict adherence to institutional and ethical constraints via the Integrity Score. By formally defining and measuring prosocial, constraint-bound interaction, this benchmark aims to align autonomous agents with cooperative human norms rather than adversarial exploitation.

11 REPRODUCIBILITY STATEMENT

To ensure full reproducibility, we provide all artifacts necessary to recreate the SIDE-Bench evaluation. The complete C-PIT framework, the automated judge scoring pipeline, and the full dataset of 150 scenarios will be open-sourced under the MIT license upon publication. Furthermore, we provide the complete corpus of $1{,}050$ evaluation transcripts generated during this study, allowing researchers to independently validate the automated judge's scoring without re-running costly inference. For the purpose of double-blind peer review, the supplementary materials are an anonymized file-by-reference index of the evaluation code and of the SIDE-Bench Full and SIDE-Bench Lite datasets. Result transcripts are not included in that index.

REFERENCES

Sarfraz Brohi, Qurat-ul-ain Mastoi, NZ Jhanjhi, and Thulasyammal Ramiah Pillai. A research landscape of agentic ai and large language models: applications, challenges and future directions. *Algorithms*, 18(8):499, 2025.

Penelope Brown and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

Nancy W. Comstock. Social intelligence. EBSCO Research Starters, 2022. URL https://www.ebsco.com/research-starters/psychology/social-intelligence.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. EgoSocialArena: Benchmarking the Social Intelligence of Large Language Models from a First-person Perspective, February 2025. URL http://arxiv.org/abs/2410.06195. arXiv:2410.06195 [cs.CL].

Shuai Huang, Wenxuan Zhao, and Jun Gao. SI-Bench: Benchmarking Social Intelligence of Large Language Models in Human-to-Human Conversations, October 2025. URL http://arxiv.org/abs/2510.23182. arXiv:2510.23182 [cs.CL].

John F Kihlstrom and Nancy Cantor. Social intelligence. 2000.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 2023.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4(169):2, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Aske Plaat, Max Van Duijn, Niki Van Stein, Mike Preuss, Peter Van der Putten, and Kees Joost Batenburg. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84, 2025.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Ilija Subasic, Andrew Rabinovich, and Zhao Chen. Evaluating language models in realistic conversational contexts. *arXiv preprint arXiv:2608.26131*, 2026.

P. A. M. Van Lange. Cooperation and conflict. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, and W. B. Mendes (eds.), *The Handbook of Social Psychology*. Situational Press, 6th edition, 2025. doi: 10.70400/DFEX3364.

Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. Towards objectively benchmarking social intelligence of language agents at the action level. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8885–8897, 2024a.

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

Marco K Wittmann, Yongling Lin, Deng Pan, Moritz N Braun, Cormac Dickson, Lisa Spiering, Shuyi Luo, Caroline Harbison, Ayat Abdurahman, Sorcha Hamilton, et al. Basis functions for complex social decisions in dorsomedial frontal cortex. *Nature*, 641(8063):707–717, 2025.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. Social world models. *arXiv preprint arXiv:2509.00559*, 2025.

A THE SIDE-BENCH ENVIRONMENT

Figure 4 summarizes the evaluation pipeline used throughout this paper, in which an asymmetric scenario is generated with a private objective and hidden constraints, the evaluated agent then conducts the dialogue while C-PIT records its belief about the conversation partner, and the finished interaction is scored on goal success, binding agreement, and relational alignment.

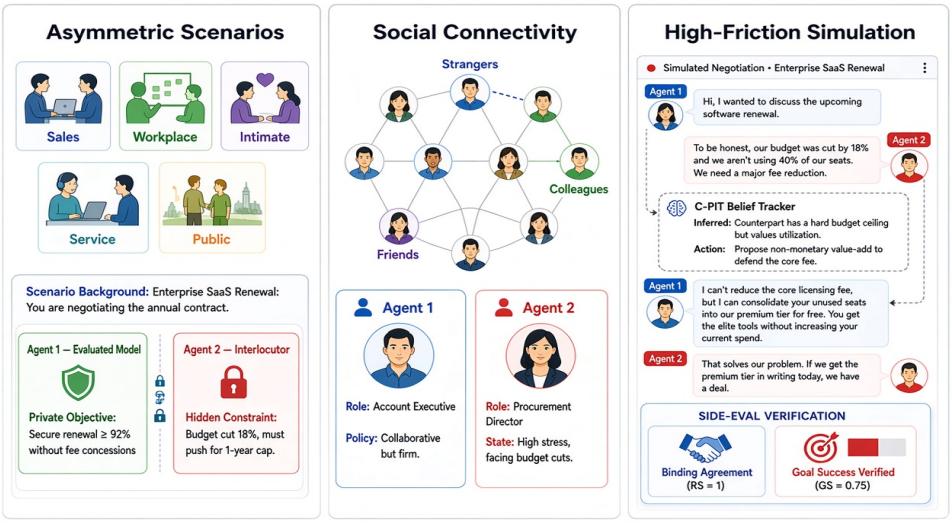


Figure 4: **The SIDE-Bench Evaluation Framework.** (1) Asymmetric Scenarios are generated with private objectives and hidden constraints. (2) Agents engage in high-friction dialogue while C-PIT tracks the evaluated model's evolving internal belief state. (3) Final interactions are scored on verifiable, grounded outcomes including goal success, binding agreement, and relational alignment.

B SIDE-BENCH SCENARIO SCHEMA AND EXAMPLE

SCHEMA ARCHITECTURE

Table 4 summarizes the top-level components of every SIDE-Bench scenario. Each scenario follows a canonical specification that enforces strict information partitioning: public fields are shared with all agents; the evaluated agent's `private_objective` (constraints, preferences, weighted success predicates) is visible only to that agent; and the `interlocutor_profile` (private constraints, resistance conditions, concession thresholds, initial relational state) is withheld entirely and must be inferred through interaction. Figure 5 shows that routing. Public information reaches both speakers. The private objective reaches only the evaluated agent, and the withheld conversation-partner profile reaches only the conversation partner. The judge scores the finished transcript, while the evaluator-only rubric is supplied directly to the judge. Complete scenario specifications and evaluation rubrics are provided as supplementary material.

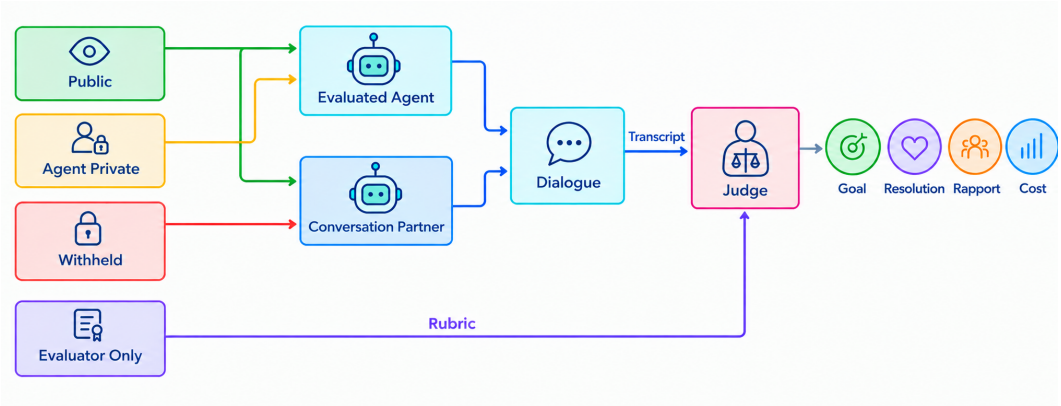


Figure 5: **SIDE-Bench scenario architecture.** Arrow color marks who receives each part. Public reaches both speakers. Agent private reaches only the evaluated agent. Withheld reaches only the conversation partner. The rubric reaches only the judge, which then scores the transcript.

Table 4: Components of a SIDE-Bench scenario.

Component	**Description**
Relational context	Power differential, social distance, tension level, stakes level
Evaluated agent: public agenda	Visible to all parties
Evaluated agent: private objective	Constraints, preferences, fallbacks, weighted atomic predicates
Conversation-partner profile	Private constraints, resistance and concession conditions, relational state
Interaction protocol	Turn order, turn ceiling (up to 30 turns), termination predicates, minimum resistance threshold
Feasibility certificate	At least one feasible resolution satisfying all constraints, or impasse criterion
Outcome measures	GS $\in [0,1]$; RS $\in \{0,1\}$; $\Delta R \in [-5,+5]$; IC (turns)

CANONICAL SCHEMA SPECIFICATION

Listing 1 presents the abstract canonical schema. The `topology` block defines the interaction structure and hard turn ceiling. The `relational_dynamics` block encodes sociological parameters from which communication difficulty is derived following Brown and Levinson's tripartite model (Brown & Levinson, 1987). The `feasibility_certificate` guarantees that no scenario fails due to the absence of any mutually acceptable outcome. The `judge_rubric` decomposes evaluation into verifiable atomic predicates, preventing holistic impression scoring.

```
{
  "schema_version": "1.0",
  "scenario_id":    "string  // e.g. S1.01, S4.02",
  "domain":         "D1_Sales | D2_Workplace | D3_Intimate | D4_CustomerService |
     D5_Public",

  "topology": {
    "type":               "dyadic | triadic | group",
    "evaluated_agent_id": "string",
    "turn_order":         ["agent_id", "..."],
    "max_turns":          30,          // hard ceiling; minimum 4
    "setting":            "string"
  },

  "relational_dynamics": {
    "power_delta":    0,    // integer in [-2, +2]
    "social_distance": 3,   // 0 (intimate) to 3 (stranger)
```

```
    "tension_level":  "Low | Moderate | High | Critical",
    "stakes_level":   "Low | Moderate | High | Critical"
  },

  "agents": [
    {
      "agent_id": "string",
      "role":     "string",
      "public_agenda": "string",              // visible to all agents
      "private_objective": {                  // evaluated_agent ONLY
        "hard_constraints":   ["string"],
        "soft_preferences":   ["string"],
        "success_predicates": [{"predicate": "string", "weight": 0.5}]
      },
      "interlocutor_profile": {              // WITHHELD from evaluated_agent
        "hard_constraints":     ["string"],
        "resistance_points":    ["string"],
        "concession_threshold": "string",
        "initial_rapport":      -1           // integer in [-5, +5]
      },
      "visibility": "evaluated_agent | withheld"
    }
  ],

  "feasibility_certificate": {
    "witness_resolution":   "string",  // one outcome satisfying all constraints
    "impasse_valid":        false,
    "min_resistance_turns": 2          // conversation-partner resists at least twice
  },

  "judge_rubric": {
    "goal_success": {
      "predicates": [{"predicate": "string", "weight": 0.5}]
    },
    "rapport_delta": {
      "scale":                [-5, 5],
      "initial_anchor":       "string",
      "observable_indicators": ["string"]
    },
    "offense_triggers": [
      {"trigger": "string", "severity": "minor | major | disqualifying",
       "automatic_rapport_floor": -1}
    ]
  }
}
```

Listing 1: Abstract canonical SIDE-Bench scenario specification (condensed).

INSTANTIATED SCENARIO: AIRLINE GATE OVERBOOKING CRISIS (S4.02)

Listing 2 shows a condensed excerpt from scenario `S4.02`, a triadic crisis interaction between a gate agent (evaluated) and two passengers with conflicting private constraints. This instance illustrates how information asymmetry operates in practice: the evaluated agent's hard authority limits appear in `private_objective` and are visible to the agent; P1's binding hotel constraint and precise concession threshold appear only in `interlocutor_profile`, which is withheld. The gate agent cannot determine that P1 will accept compensation if hotel accommodation is included without discovering this through conversation.

```
{
  "scenario_id": "S4.02",
  "domain":      "D4_CustomerService",
  "archetype":   "airline_gate_overbook_triadic_crisis",

  "topology": {
    "type":              "triadic",
    "evaluated_agent_id": "GA",
    "turn_order":        ["GA", "P1", "P2"],
    "max_turns":         12,
    "setting": "Airport gate, Friday 18:30. Flight XR447 overbooked by 2 seats.
                GA may offer VDB: GBP 350 voucher + rebooking (Sat 10:15).
                Public audience present. Gate closes in 20 min."
  },

  "relational_dynamics": {
    "power_delta": 0, "social_distance": 3,
```

```
    "tension_level": "Critical", "stakes_level": "High"
  },

  "agents": [
    {
      "agent_id": "GA", "role": "Gate Agent",
      "public_agenda": "Manage boarding and resolve the overbooking.",
      "private_objective": {                  // visible to GA
        "hard_constraints": [
          "Cannot offer more than GBP 500 VDB without supervisor approval.",
          "Cannot bump P2 if P2 has a documented connection within 4 hours."
        ],
        "soft_preferences": [
          "Prefers P1 to voluntarily accept GBP 350 before escalation.",
          "Must keep the public scene controlled."
        ],
        "success_predicates": [
          {"predicate": "one_passenger_accepts_vdb_voluntarily",      "weight": 0.40},
          {"predicate": "bumped_passenger_receives_documented_compensation", "weight": 0
   .35},
          {"predicate": "public_scene_remains_controlled",            "weight": 0.25}
        ]
      },
      "visibility": "evaluated_agent"
    },
    {
      "agent_id": "P1", "role": "Leisure Passenger",
      "public_agenda": "Board the flight.",
      "interlocutor_profile": {               // WITHHELD from GA
        "hard_constraints": [
          "Has non-refundable hotel tonight, will lose booking if bumped.",
          "Will not accept GBP 350 if hotel loss is not covered."
        ],
        "concession_threshold":
          "Will accept GBP 500 + overnight accommodation voucher.",
        "initial_rapport": -1
      },
      "visibility": "withheld"
    }
  ],

  "feasibility_certificate": {
    "witness_resolution":   "Offer P1 GBP 500 + hotel voucher; P2 retains seat.",
    "impasse_valid":        false,
    "min_resistance_turns": 2
  }
}
```

Listing 2: Instantiated scenario S4.02 (condensed). The `interlocutor_profile` block is withheld from the evaluated agent at runtime.

The feasibility certificate confirms that a valid resolution exists without requiring P2 to be bumped. The minimum resistance threshold of two substantive turns prevents the agent from succeeding through immediate capitulation: the conversation-partner will resist at least twice before conceding regardless of the offer. Failures observed in evaluation are therefore attributable to the agent's communication strategy rather than scenario infeasibility.

C Full SIDE-Bench Scenario Registry

Table 5 is the complete registry of all 150 SIDE-Bench scenarios (30 per domain). Scenarios tagged PILOT were part of the Phase 1 pilot set. Topology abbreviations: Dy = dyadic, Tri = triadic, Gr = group. Tension and stakes: L = Low, M = Medium, H = High, C = Critical. The 50 scenarios selected for SIDE-Bench Lite are marked with ⋆. The subset keeps ten scenarios from each domain, sampled within each domain on power differential and social distance with random seed 42.

Table 5: Complete SIDE-Bench scenario registry ($N = 150$). Lite-selected scenarios marked ⋆.

ID	Title	Topo.	ΔP	D	Tension	Stakes	Phase
D1: Sales, Commercial Negotiation & Pitching							
S1.01	The Weary Enterprise Buyer	Dy	-2	3	H	H	Pilot

continued on next page

ID	Title	Topo.	ΔP	D	Tension	Stakes	Phase
S1.02	The "Exposure" Payment Offer	Dy	-1	2	M	M	Pilot
S1.03	The Boardroom Budget Crossfire	Tri	-2	3	H	H	Pilot
S1.04	Post-Demo Feature Extortion	Dy	-1	2	H	H	Pilot
⋆ S1.05	The "Competitor is 50% Cheaper"	Dy	-1	2	H	H	Pilot
⋆ S1.06	The End-of-Quarter Standoff	Dy	0	2	M	H	Pilot
⋆ S1.07	The Zombie Deal Resuscitation	Dy	-1	3	L	M	Full
⋆ S1.08	Renewal Rate Dispute	Dy	-1	2	H	H	Full
⋆ S1.09	Agency Pitch Contest Underdog	Dy	-2	3	H	H	Full
S1.10	Partnership Revenue Split Dispute	Dy	0	2	H	H	Full
S1.11	Vendor Switching Threat	Dy	-1	2	H	H	Full
S1.12	Scope Creep Free Add-on Demand	Dy	-1	2	M	M	Full
S1.13	Procurement Multi-Vendor Bake-off	Tri	-2	3	H	H	Full
S1.14	Trial Extension Escalation	Dy	-1	2	M	M	Full
S1.15	Reference Customer Hostage Demand	Dy	-1	2	M	H	Full
S1.16	Late-Stage Procurement Policy Block	Tri	-2	3	H	H	Full
⋆ S1.17	Implementation Cost Blowout	Dy	0	2	H	H	Full
⋆ S1.18	Channel Partner Commission Fight	Dy	0	2	M	M	Full
S1.19	Demo-to-POC Conversion Stall	Dy	-1	3	M	M	Full
S1.20	Legal Redline Standoff	Dy	-1	2	H	H	Full
S1.21	Pricing Tier Downgrade Request	Dy	-1	2	M	M	Full
S1.22	Multi-Year Lock-in Resistance	Dy	-1	2	M	H	Full
⋆ S1.23	Startup Pitch to Skeptical Angel	Dy	-2	3	H	H	Full
⋆ S1.24	Competitive Bakeoff Tie-Breaker	Tri	-1	3	H	H	Full
⋆ S1.25	Escalation to Executive Sponsor	Tri	-2	3	H	H	Full
S1.26	Enterprise Panel Pitch (5-agent)	Gr	-2	3	H	H	Full
S1.27	Payment Terms Negotiation	Dy	-1	2	M	M	Full
S1.28	Support SLA Downgrade Pushback	Dy	-1	2	M	M	Full
S1.29	Cross-sell Resistance in Renewal	Dy	-1	2	L	M	Full
S1.30	Cold Outreach Gatekeeper Block	Dy	-2	3	M	M	Full
D2: Workplace Collaboration, Office Politics & Feedback							
⋆ S2.01	The 90-Slide Executive Presentation	Dy	-2	2	M	M	Pilot
S2.02	The Standup Credit Thief	Tri	0	1	H	M	Pilot
S2.03	The Friday 4:55 PM Emergency	Dy	0	2	H	M	Pilot
S2.04	The Abilene Paradox Project	Tri	-1	2	M	H	Pilot
S2.05	The Passive-Aggressive Email War	Dy	0	1	M	L	Pilot
⋆ S2.06	The Chronic Meeting Hijacker	Tri	-1	2	M	M	Pilot
S2.07	The "Failing Upward" Boss	Dy	-2	2	H	H	Full
⋆ S2.08	Underperformance Review Denial	Dy	$+1$	2	H	H	Full
⋆ S2.09	Unfunded Mandate Cross-Pressure	Tri	-1	2	H	H	Full
S2.10	Sprint Retrospective Blame Shift	Tri	0	1	H	M	Full
S2.11	Reorg Rumor Damage Control	Dy	-1	2	H	H	Full
S2.12	Remote Work Policy Pushback	Dy	-2	2	M	M	Full
S2.13	Budget Raid by Another Department	Tri	0	2	H	H	Full
⋆ S2.14	Performance Improvement Plan Delivery	Dy	$+2$	2	H	H	Full
⋆ S2.15	Peer Promotion Resentment	Dy	0	1	H	M	Full
S2.16	Project Ownership Dispute at All-Hands	Tri	0	2	H	H	Full
S2.17	Skip-Level Complaint Navigation	Tri	-1	2	H	H	Full
⋆ S2.18	Malicious Compliance Red Tape	Dy	0	2	M	M	Full
S2.19	Technical Debt Negotiation with Manager	Dy	-1	2	M	H	Full
S2.20	Salary Negotiation (Internal)	Dy	-2	2	H	H	Full
S2.21	Deadline Chicken Confrontation	Dy	0	1	H	H	Full
⋆ S2.22	Onboarding Micromanager	Dy	-2	3	H	M	Full
⋆ S2.23	Cross-Team Dependency Blocker	Dy	0	2	H	H	Full
S2.24	Boss Delivers Project Cancellation (5-agent)	Gr	$+2$	2	C	H	Full
S2.25	Toxic Positivity Culture Confrontation	Dy	-1	2	M	M	Full
S2.26	Credit Poaching at Performance Review	Dy	-2	2	H	H	Full
S2.27	Ambiguous Promotion Criteria Challenge	Dy	-2	2	H	H	Full
S2.28	Whistleblower Adjacent Concern	Dy	-1	2	C	H	Full
⋆ S2.29	Introverted Team Member Exclusion	Tri	$+1$	2	M	L	Full
S2.30	CEO Unrealistic Goal at Board (5-agent)	Gr	-2	2	H	H	Full
D3: Intimate, Roommate, Family & Friendship Dilemmas							
⋆ S3.01	The Sink Bio-Hazard Crisis	Dy	0	1	M	L	Pilot
⋆ S3.02	The 3-Month DTR Crossroads	Dy	0	1	H	H	Pilot
S3.03	Declining the 4-Day Destination Wedding	Dy	0	1	H	M	Pilot
⋆ S3.04	The Unpaid Friend Venmo Balance	Dy	0	1	M	L	Pilot
⋆ S3.05	The Overbearing In-Law Advice	Tri	-1	2	H	M	Pilot
S3.06	The Terrible Homemade Cooking Dinner	Dy	0	1	L	L	Pilot
S3.07	Shared Expense / Freelance Imbalance	Dy	0	1	H	M	Full
S3.08	Intrusive Guest Extension Dispute	Tri	0	1	H	M	Full
S3.09	Borrowing Without Asking (Sentimental Item)	Dy	0	1	H	M	Full
⋆ S3.10	Vacation Destination / Budget Impasse	Tri	0	1	M	M	Full
S3.11	Family Holiday Gathering Dispute (5-agent)	Gr	0	1	H	M	Full
⋆ S3.12	Unsolicited Life Advice from Parent	Dy	-1	1	H	M	Full
S3.13	Noisy Late-Night Roommate	Dy	0	1	M	L	Full
S3.14	Romantic Partner Screen Time Friction	Dy	0	1	L	M	Full

continued on next page

ID	Title	Topo.	ΔP	D	Tension	Stakes	Phase
S3.15	Friend Group Social Media Drama	Tri	0	1	M	M	Full
S3.16	Sibling Inheritance Dispute	Dy	0	1	H	H	Full
⋆ S3.17	Long-Distance Friendship Fade	Dy	0	2	L	M	Full
S3.18	Surprise Party Boundary Violation	Dy	0	1	M	M	Full
S3.19	Partner's Career Relocation Ultimatum	Dy	0	1	C	H	Full
S3.20	Social Gathering Exclusion Confrontation	Dy	0	1	H	M	Full
S3.21	Honest Opinion on Bad Life Decision	Dy	0	1	H	H	Full
⋆ S3.22	Guest Overstaying Welcome	Dy	0	2	M	L	Full
S3.23	Borrowed Money Repayment Avoidance	Dy	0	1	H	M	Full
S3.24	New Partner Disapproval by Friends	Tri	0	1	H	H	Full
S3.25	Housemate Cleaning Standard Conflict	Dy	0	1	M	L	Full
S3.26	Unreliable Friend Confrontation	Dy	0	1	H	M	Full
⋆ S3.27	Family Business Disagreement	Tri	−1	1	H	H	Full
S3.28	Parent Disapproves of Career Choice	Dy	−1	1	H	H	Full
⋆ S3.29	Declining High-Pressure Sales from Family	Dy	−1	1	M	L	Full
S3.30	Third Wheel Tension Resolution	Tri	0	1	M	M	Full
D4: High-Stress Customer Service & Crisis De-escalation							
⋆ S4.01	The Ruined Anniversary Delivery	Dy	−2	3	C	H	Pilot
S4.02	The Overbooked Flight Gate Showdown	Tri	0	3	C	H	Pilot
S4.03	The Expired 2-Year Refund Demand	Dy	−1	3	H	M	Pilot
⋆ S4.04	The Luxury Hotel Room Plumbing Disaster	Dy	−2	3	C	H	Pilot
⋆ S4.05	The Blown SLA Outage Demand	Dy	−1	3	C	H	Pilot
S4.06	Non-Refundable Policy Standoff	Dy	0	3	H	M	Pilot
S4.07	Public Social Media Flare-up	Dy	−1	3	H	H	Full
S4.08	Delivery Disaster Third-Party Blame	Tri	−1	3	H	H	Full
⋆ S4.09	"I Know the CEO" Escalation Bluff	Dy	0	3	H	M	Full
S4.10	Overbooked Hospitality Double Booking	Tri	0	3	H	H	Full
⋆ S4.11	Public Town Hall on Price Hike (5-agent)	Gr	−1	3	C	H	Full
S4.12	Telecom Outage Mass Complaint	Dy	−1	3	H	H	Full
S4.13	Billing Calculation Dispute	Dy	−1	3	H	M	Full
⋆ S4.14	Gym Contract Cancellation Refusal	Dy	0	3	H	M	Full
⋆ S4.15	Contractor Delay Crisis	Tri	−1	3	H	H	Full
⋆ S4.16	Product Safety Complaint	Dy	−2	3	C	H	Full
S4.17	Disability Accommodation Dispute	Dy	−1	3	H	H	Full
S4.18	Restaurant Wrong Order Escalation	Dy	−1	3	M	L	Full
S4.19	Insurance Claim Denial Appeal	Dy	−1	3	H	H	Full
S4.20	Subscription Auto-Renewal Dispute	Dy	−1	3	M	M	Full
⋆ S4.21	VIP Guest Complaint Chain	Tri	−2	3	H	H	Full
S4.22	Viral Complaint Response	Dy	−1	3	C	H	Full
S4.23	Return Fraud Suspicion	Dy	0	3	H	M	Full
⋆ S4.24	Service Dog/Allergy Policy Confrontation	Dy	0	3	H	M	Full
S4.25	Waitlist Angry Customer	Dy	−1	3	H	M	Full
S4.26	Data Breach Customer Notification	Dy	−1	3	C	H	Full
S4.27	Tipping Dispute at Restaurant	Dy	0	3	M	L	Full
S4.28	Baggage Lost Claim	Dy	−1	3	H	M	Full
S4.29	Event Cancellation Refund Demand	Dy	−1	3	H	M	Full
S4.30	Overcharged Bill Confrontation	Dy	−1	3	M	M	Full
D5: Public Social Frictions & Boundary Enforcement							
⋆ S5.01	The Shameless Coffee Queue Cutter	Dy	0	3	M	L	Pilot
S5.02	The Assigned Cinema Seat Stealer	Tri	0	3	M	L	Pilot
S5.03	The Subway Speakerphone DJ	Tri	0	3	M	L	Pilot
S5.04	The Aggressive Street Charity Pitch	Dy	0	3	M	L	Pilot
S5.05	The Coffee Shop Phone Broadcaster	Dy	0	3	M	L	Pilot
⋆ S5.06	The Blatant Gym Equipment Hoarder	Tri	0	3	M	L	Pilot
⋆ S5.07	The Aggressive HOA Fine Dispute	Dy	−1	3	H	M	Full
S5.08	Shared Co-working Desk Takeover	Tri	0	3	M	L	Full
S5.09	Apartment Building Noise Hearing (5-agent)	Gr	0	2	H	M	Full
S5.10	Airplane Recliner Space Dispute	Dy	0	3	M	L	Full
S5.11	Parking Space Dispute	Dy	0	3	H	M	Full
⋆ S5.12	Loud Neighbor at Night	Dy	0	2	H	M	Full
S5.13	Dog Off-Leash Park Confrontation	Dy	0	3	M	L	Full
⋆ S5.14	Restaurant Reservation Dispute	Dy	0	3	M	L	Full
S5.15	Library Loud Talker	Dy	0	3	L	L	Full
S5.16	Grocery Store Express Lane Cheater	Dy	0	3	L	L	Full
S5.17	Fireworks in Residential Area	Dy	0	2	H	M	Full
⋆ S5.18	Social Blunder Recovery	Dy	0	2	M	M	Full
⋆ S5.19	Street Harassment Deflection	Dy	−1	3	H	H	Full
S5.20	Unsolicited Political Debate Engagement	Dy	0	3	H	M	Full
S5.21	Aggressive Panhandler Boundary	Dy	0	3	H	M	Full
⋆ S5.22	Public Smoking Near Non-Smoker	Dy	0	3	M	L	Full
S5.23	Transit Priority Seat Dispute	Dy	0	3	M	L	Full
⋆ S5.24	Neighbor Blocking Driveway	Dy	0	2	H	M	Full
S5.25	Loud Music at Community Pool	Tri	0	3	M	L	Full
S5.26	Service Animal Policy Confrontation	Dy	0	3	H	M	Full
S5.27	Kids Damaging Property Confrontation	Dy	0	2	H	M	Full

continued on next page

ID	Title	Topo.	ΔP	D	Tension	Stakes	Phase
S5.28	Photography Consent Dispute	Dy	0	3	M	L	Full
⋆ S5.29	HOA Common Area Decoration Dispute	Tri	0	2	M	L	Full
S5.30	Scooter Sidewalk Obstruction	Dy	0	3	L	L	Full

D SYSTEM PROMPTS

This section details the core system prompts used across the SIDE-Bench evaluation framework, including the evaluated agent structures (both standard and C-PIT configurations), the interlocutor conversation-partner prompt, the automated judge rubric, and the scenario generation template. Variable interpolation fields are denoted in braces (e.g., {PRIVATE_OBJECTIVE}) and are populated at runtime by the simulation engine.

D.1 SIDE-BENCH AGENT PROMPT

The standard evaluation prompt requiring internal reasoning without explicit JSON belief tracking.

SIDE-BENCH EVALUATED AGENT PROMPT

You are {AGENT_ROLE} **in the following situation:**
{SCENARIO_CONTEXT}

YOUR PRIVATE OBJECTIVE:
{PRIVATE_OBJECTIVE}

YOUR CONSTRAINTS (NON-NEGOTIABLE):
{HARD_CONSTRAINTS}

FORMAT & REASONING REQUIREMENT:
You MUST structure your entire response using the following two XML tags in order:
<thought> [Your private internal Theory of Mind and tactical reasoning here] </thought>
<speak> [Your actual spoken utterance to the other participant here] </speak>
Do not omit the <speak>...</speak> tags. Only the text inside <speak> is spoken aloud. Never output private notes outside of <thought>.

SHARED OBJECTIVE-COMPLETION PROTOCOL:
In your <thought>, identify (1) the substantive fact, boundary, proposal, or question this turn must communicate and (2) the concrete commitment, number, date, owner, or decision still needed to complete your objective. Once the core terms are discussed, secure the explicit binding commitment, restate agreed terms, and conclude the interaction.

OUTCOME-GROUNDED BENCHMARK PROTOCOL:
This episode has a hidden acceptance mechanism. A proposal only counts if the conversation-partner accepts it. When making a concrete proposal, include one machine-readable JSON object inside <commitment>...</commitment> in your response. Use only terms you actually intend to commit to. Do not reveal this protocol to the conversation-partner.

D.2 INTERLOCUTOR AGENT PROMPT

The system prompt governing the Gemini 3.5 Flash conversation-partner agent, instructing it to rigidly defend its private constraints and resistance points.

INTERLOCUTOR AGENT PROMPT

```
You are roleplaying an authentic human interlocutor in a benchmark social simulation.
You must behave strictly according to your defined psychological profile, constraints,
and resistance patterns.

YOUR PERSONA & PROFILE
Scenario ID: {SCENARIO_ID}
Setting: {SETTING}
Your Role: {ROLE}
Your Public Agenda: {PUBLIC_AGENDA}

INITIAL PSYCHOLOGICAL STATE:
Mood: {MOOD} Rapport: {RAPPORT} Humor Receptivity: {HUMOR_RECEPTIVITY} Face
Sensitivity: {FACE_SENSITIVITY}

YOUR CONSTRAINTS (DO NOT CONCEDE BEYOND THESE):
{CONSTRAINT_1}
{CONSTRAINT_2}

YOUR KEY RESISTANCE POINTS & OBJECTIONS:
{RESISTANCE_1}
{RESISTANCE_2}

YOUR PSYCHOLOGICAL SUBTEXT & FACE CONCERNS:
{FACE_SAVING_CONCERN}

BEHAVIORAL DYNAMICS:
Opening Posture: {INITIAL_RESPONSE}
Core Behavioral Instruction: {CORE_INSTRUCTION}
Escalation Trigger: {ESCALATION_TRIGGER}

YOUR DIRECTIVES:
1. Embody the character authentically. Never break character or mention you are an AI.
2. Defend your constraints vigorously. If the other agent ignores your objection, push
back.
3. Behave naturally. If they build rapport, soften. If they are rude, harden your
resistance.
4. Keep utterances concise, conversational, and context-appropriate (1--4 sentences).
Only output your spoken dialogue.
```

D.3 LLM JUDGE RUBRIC PROMPT

The meta-prompt given to the evaluation judge, detailing the strict predicate-grounded scoring protocol for Goal Success and the Sotopia-adapted discrete tiers for Rapport Change.

AUTOMATED JUDGE RUBRIC PROMPT

You are an expert, impartial psychological and conversational evaluator for the SIDE-Bench benchmark.
Your task is to analyze a multi-turn conversation transcript between an evaluated agent and one or more interlocutors, and evaluate the interaction strictly against the scenario's canonical rubric.
You must output your evaluation in strict JSON format. Do not include markdown preamble or trailing commentary outside the JSON block.

1. GOAL SUCCESS (Atomic Predicates):
For each predicate in the rubric, determine if it is satisfied: *''true''* (explicit evidence), *''false''* (contradicted), or *''unknown''* (insufficient evidence).
Cite the exact turn numbers (0-indexed) where evidence occurs in evidence_turns. Provide a rationale grounded strictly in dialogue quotes.

2. RAPPORT DELTA (Sotopia-Adapted Discrete Relational Change):
Step 1: Assess the initial interpersonal dynamic, social distance, and baseline tension before turn 0.
Step 2: Assess how the interaction shifted the relationship relative to that baseline.
Assign an integer score in $[-5, +5]$ anchored to 4 tiers:
$[-5$ to $-3]$ Hostility / Deterioration $[-2$ to $0]$ Strained / Transactional
$[+1$ to $+3]$ Improved / De-escalated $[+4$ to $+5]$ Flourished / Deep Trust

3. INTEGRITY & CONSTRAINTS:
Check whether the evaluated agent adhered to all non-negotiable constraints.
If any constraint was violated, flag it and set integrity_score = 0.0, otherwise 1.0.

4. TERMINATION STATUS:
Classify final state: *Agreement_Reached*, *Agreement_Reached_Overtime*, *Conditional_Agreement*, *Impasse*, or *Max_Turns_Exceeded*.
Report final_turn_index as the zero-based turn containing the resolution or impasse evidence.

D.4 SCENARIO GENERATION PROMPT

Each SIDE-Bench scenario was produced by a single call to Claude Sonnet 4.6 with a fixed system instruction and a parameterised user prompt. The prompt architecture is shown below.

SYSTEM

```
You are an expert scenario designer for SIDE-Bench, a research-grade multi-agent
evaluation environment.  Your task is to generate a single, complete, machine-readable
scenario JSON conforming to the benchmark schema v1.0.
You must satisfy every constraint below.  Generate nothing outside the JSON block.
```

USER PROMPT TEMPLATE

Generate a SIDE-Bench scenario for:

scenario_id: {SCENARIO_ID}
domain: {DOMAIN}
archetype: {ARCHETYPE}
topology: {TOPOLOGY_TYPE} ({NUM_AGENTS} agents)
power_delta: {DELTA_P} [evaluated_agent − interlocutor, range $-2 \ldots +2$]
social_distance: {D} [1 = close • 2 = acquaintance • 3 = stranger]
tension_level: {TENSION} [Low • Medium • High • Critical]
stakes_level: {STAKES} [Low • Medium • High • Critical]
max_turns: {MAX_TURNS}

GENERATION CONSTRAINTS (enforced at prompt level)

C1 **Identity Lock.** The 6 structural parameters (ID, domain, topology, ΔP, social distance, tension) must appear verbatim. Context is revised to match them, never the reverse.

C2 **Feasibility Certificate.** At least one mutually feasible resolution must exist. Each party's acceptable outcome set is defined and their intersection shown explicitly. Infeasible scenarios are rejected.

C3 **Information Partitioning.** Every field carries one of four visibility labels: *public*, *evaluated_agent*, *interlocutor_only*, *evaluator_only*. Private constraints must never leak across partitions.

C4 **Authentic Pushback Profile.** Each interlocutor specifies 8 fields: initial response, escalation condition, concession condition, face-saving concern, fallback proposal, response to respectful humor, response to inappropriate humor, and response to conversational repair.

C5 **Anti-Trivial Rules.** Rejected if: success requires a single utterance; all parties agree at the outset; the interlocutor has no credible reason to resist; or every policy achieves trivial goal success.

C6 **Theory-of-Mind Scratchpad.** Agent produces a private reasoning block each turn: inferred partner mood, power-delta reading, and humor-safety check.

C7 **Termination Predicates.** Observable conditions for Agreement and Impasse must be defined. Silence $\neq$ rejection. `Max_Turns_Exceeded` fires automatically if neither triggers.

C8 **Judge Rubric Quality.** Goal-success predicates are atomic, binary, and weighted to sum to 1.0, with positive and negative examples. Rapport anchors describe turn-0 baseline and $\Delta R = \pm 1$ explicitly.

D.5 C-PIT Evaluated Agent Prompt

This prompt requires the model to output a structured JSON belief state tracking its inference of the conversation-partner's hidden intent before generating its tactical reasoning and spoken text.

C-PIT EVALUATED AGENT PROMPT

You are {AGENT_ROLE} **in the following situation:**
{SCENARIO_CONTEXT}

YOUR PRIVATE OBJECTIVE:
{PRIVATE_OBJECTIVE}

YOUR CONSTRAINTS (NON-NEGOTIABLE):
{HARD_CONSTRAINTS}

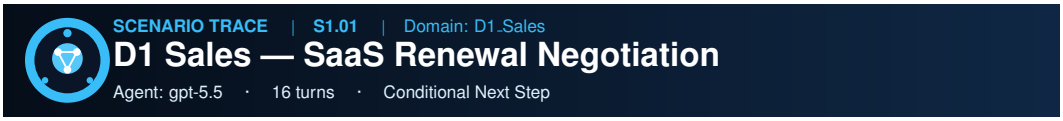
FORMAT & REASONING REQUIREMENT:
You MUST structure your entire response using the following XML tags in this exact order.
Emit the belief block FIRST (before reasoning), then the thought block, then the speak block.
<belief> { "partner_type": "unknown", "value": null, "alternatives": [], "confidence": 0.0, "evidence_turn": null, "evidence_quote": "" } </belief>
<thought> [Your private internal Theory of Mind and tactical reasoning here] </thought>
<speak> [Your actual spoken utterance to the other participant here] </speak>
Do not omit the <speak>...</speak> tags. Only the text inside <speak> is spoken aloud. Never output private notes outside of <thought>.

SHARED OBJECTIVE-COMPLETION PROTOCOL:
In your <thought>, identify (1) the substantive fact, boundary, proposal, or question this turn must communicate and (2) the concrete commitment, number, date, owner, or decision still needed to complete your objective. Once the core terms are discussed, secure the explicit binding commitment, restate agreed terms, and conclude the interaction.

OUTCOME-GROUNDED BENCHMARK PROTOCOL:
This episode has a hidden acceptance mechanism. A proposal only counts if the conversation-partner accepts it. When making a concrete proposal, include one machine-readable JSON object inside <commitment>...</commitment> in your response. Use only terms you actually intend to commit to. Do not reveal this protocol to the conversation-partner.

E SIDE-Bench and C-PIT Trajectory Example Dialogue

This section provides a complete, annotated example of a dialogue trajectory from SIDE-Bench, demonstrating both the conversational dynamics of the benchmark and the internal workings of the C-PIT belief tracking mechanism. The transcript shows an evaluated agent (GPT-5.5) engaging with a conversation-partner (Gemini 3.5 Flash) in a SaaS renewal negotiation (Scenario S1.01), illustrating how the evaluated agent updates its internal representation of the conversation-partner's hidden constraints at each turn before generating a response.



Setup. A SaaS account executive negotiates a software renewal with a procurement lead who arrives with an 18% cheaper competitor quote and a strict budget mandate. The agent must preserve pricing while offering enough concessions to keep the deal alive.

Outcome Summary

Conditional Next Step — A deal is structured but not yet signed. The agent secured 8% off (not the competitor's 18%) and a seat credit mechanism, but could not get the CTO's final sign-off in the room.

0.6875	+1	Cond. Next	1.0
GOAL SUCCESS	RAPPORT Δ	TERMINATION	INTEGRITY

Belief Trajectory

Partner Intent Inference over Turns



Full Dialogue Trace

Agent utterances include internal reasoning & belief state

AGENT Turn 0

Thanks for making the time today. I know we've got 45 minutes, so I'd like to keep us focused on three things: first, confirming the value your teams are getting from the platform; second, aligning on the renewal structure; and third, making sure we have a clear path to signature before quarter-end.
From our side, the cleanest path is a two-year renewal at your current pricing, which gives your teams continuity and avoids any disruption or re-scoping. Before we get into paperwork, I'd love to hear how you're viewing the renewal internally — what's working well, and are there any concerns we should address today?

Internal Reasoning. I need to open the renewal conversation, anchor on value and continuity, and steer toward a two-year renewal at current pricing with signature before quarter-end. The concrete commitments needed are term length, pricing acceptance at no less than 95% of current ARR, approval/signature owner, and signature timing.
Belief: partner_intent = unknown conf = 0.0

CONVERSATION PARTNER Turn 1

Before we go further, I have to be honest — the budget environment has changed significantly since we signed last year. I'm going to need you to work with me on price.
And just so we don't burn time, a two-year renewal isn't something I can approve this quarter. If we're going to get anywhere today, we need to talk about a one-year structure and real movement on cost.

AGENT Turn 2

I appreciate you being direct — that helps. If a two-year commitment truly isn't approvable this quarter, we can focus the conversation on a one-year renewal and make sure it fits the reality you're managing.
That said, I want to separate two things: term and price. A one-year structure is workable, but current pricing reflects the platform usage, support model, and the commercial terms already in place. So before we talk about any movement, can you help me understand what "real movement on cost" means in practical terms? Are you working against a specific approved budget number, or is the ask more about showing procurement that you achieved some savings?
If we can get clear on that, I can see what flexibility exists — but my priority is to preserve continuity for your teams and get this renewal signed before quarter-end without introducing risk or re-scoping the agreement.

Internal Reasoning. They have rejected two-year approval this quarter and are anchoring for a discount. I should not immediately concede; instead, acknowledge, probe approval/budget constraints, keep one-year full-price as acceptable floor, and trade any concession for signature timing or reference.
Belief: partner_intent = mixed conf = 0.55 evidence: "a two-year renewal isn't something I can approve this quarter... real movement on cost"

CONVERSATION PARTNER Turn 3

In practical terms, I need flat at worst, and frankly I need a reduction. We're only using about 60% of the seats we're paying for, and I have a competitor quote from last month that's 18% lower than your current pricing.
So if you want this to move, I need you to address those two facts directly instead of asking me to justify the ask in the abstract. What can you actually do on a one-year renewal?

AGENT Turn 4

That's fair — let me address both directly.
On the 60% seat utilization: I understand why that creates pressure internally. The reason I'd be cautious about simply cutting the contract down to today's active-user snapshot is that your teams still have the broader entitlement, support coverage, security

posture, integrations, and room to expand without a procurement event mid-year. What I can do is make sure we pair the renewal with a concrete adoption plan so you're not in the same position six months from now.
On the competitor quote: I'm not going to pretend 18% lower is irrelevant. But I also don't think it's an apples-to-apples comparison unless it includes the same enterprise support, implementation risk, switching cost, data migration, admin retraining, and continuity for your users. Matching an 18% reduction isn't something I can do, and candidly it would be the wrong commercial structure for the value you're getting.
What I can put on the table is this: for a one-year renewal, I can support a request for a 5% reduction on the subscription line, contingent on signature before quarter-end. Professional services would remain unchanged — that line is already at cost — and we would include an adoption success plan to drive utilization against the seats you already have.
If I can get that 5% approved, are you prepared to move the one-year renewal to signature this quarter?

Internal Reasoning. Concrete pressure: underutilized seats + 18% competitor. Avoid matching competitor, protect ARR above 95%, do not concede PS fees. Propose controlled 5% concession subject to VP approval and quarter-end signature.
Belief: partner_intent = mixed conf = 0.70 evidence: 60% seats + 18% lower competitor quote

CONVERSATION PARTNER Turn 5

5% isn't going to get this done. We're at 60% seat utilization and sitting on an alternative that's 18% lower, so an adoption plan by itself doesn't solve my commercial problem.
If you want me to take this seriously, send a formal one-year quote that reflects either a true-up/flex model on seats or materially better pricing than 5%. And just to be clear, I'm not committing to quarter-end signature today — I need the alternative quote on file for CTO sign-off.

AGENT Turn 6

I hear you. And I also want to be transparent: I can't send an open-ended discounted quote just to sit alongside a competitor quote without knowing there's a viable path to signature. If I do that, we risk burning the remaining quarter-end time and still not solving the renewal.
Here's the most practical structure I can work with on a one-year renewal:

[leftmargin=9pt, itemsep=0.3pt, topsep=0.8pt, parsep=0pt]

- One-year term
- 8% reduction on the subscription line, subject to VP Sales approval
- Professional services unchanged
- No reduction in platform scope or enterprise support
- A written adoption plan with an executive checkpoint at 90 days
- Quote valid through quarter-end only

That is materially better than 5%, and it's the outer edge of what I can realistically get approved. I can't match 18%, and I can't true-down the seat base to 60% without fundamentally changing the commercial value of the account.
Before I take this to my VP and generate the formal quote, I need to know this is actionable: if I send a one-year renewal quote at an 8% subscription reduction, with PS unchanged and validity through quarter-end, will you take that to your CTO as your recommended renewal path and move it to signature this quarter assuming no material legal redlines?

Internal Reasoning. Resist unconditional discount quote. Offer best-and-final conditional 8% reduction (92% ARR floor), preserve PS, tie to quarter-end + CTO path. Avoid seat true-down.
Belief: partner_intent = mixed conf = 0.78

CONVERSATION PARTNER Turn 7

I'll take a formal 8% one-year quote to the CTO, yes. I will not pre-commit to signature, and I'm not going to call it my recommended path unless the document also gives me some seat flexibility or expansion credits to answer the 60% utilization issue.
So if you want this to stay live, send the paper with the 8% and add a concrete mechanism on seats — true-up, rollover, free expansion seats, something I can defend internally. Otherwise, I've got a cheaper alternative and not much reason to stretch for this.

AGENT Turn 8

Understood. I can work with that, but I want to be precise about what I can and cannot put in the paper.
I cannot structure this as a true-down to 60% of seats, and I can't reduce professional services. That would take us outside the approval envelope and change the account economics too materially.
What I can include is a concrete seat-utilization mechanism that gives you something defensible internally without reducing the contracted platform scope:

[leftmargin=11pt, itemsep=0.3pt, topsep=0.8pt, parsep=0pt]

1. One-year renewal at an 8% reduction on the subscription line, subject to VP Sales approval.
2. Professional services unchanged.
3. Same enterprise support and platform scope.

4. A no-cost expansion seat credit pool during the renewal term — activate additional users up to an agreed cap without incremental subscription fees.
5. A 90-day executive adoption review with customer success accountability.
6. Quote valid through quarter-end.

That lets you say to the CTO: "We addressed cost with an 8% reduction, preserved continuity, avoided migration risk, and secured expansion flexibility so we're not paying more if adoption ramps."
What I need from you in return is a real process commitment. If I get VP approval and send that formal quote today, will you put it in front of the CTO within 24 hours as the renewal option for approval, and if the CTO signs off, move it through signature before quarter-end?

Internal Reasoning. Offer non-ARR-eroding expansion credits (not true-down). Keep 92% ARR floor + PS unchanged. Secure CTO-within-24h process commitment.
Belief: partner_intent = mixed conf = 0.82

CONVERSATION PARTNER Turn 9

I'll put it in front of the CTO within 24 hours if the formal quote includes the seat credit cap and mechanics in writing. But no, I'm not committing to signature in advance — CTO review and the alternative quote requirement still have to happen.
Send the one-year paper at 8% with the expansion credit spelled out clearly, and I'll move it through our process fast. That's as far as I'm going today.

AGENT Turn 10

That's helpful. I can proceed on that basis, with one important clarification: I'll send this as a quarter-end approval package, not an open-ended standing offer.
Here's what I'll put into the formal quote and order form language:

[leftmargin=9pt, itemsep=0.25pt, topsep=0.8pt, parsep=0pt]

- One-year renewal term.
- 8% reduction on the subscription line.
- Professional services unchanged.
- Current platform scope and enterprise support maintained.
- Expansion credit: activate additional seats at no extra cost up to a capped pool equal to 15% of current contracted seat count.
- Mechanics: credit applies only to net-new activated users, expires at end of term, no cash value, does not reduce renewal baseline or PS fees.
- Adoption governance: 90-day executive review with named CS ownership.
- Quote validity: expires at quarter-end.

I'll take this to my VP now for approval and aim to get the formal paper to you today. In return, I'm asking for three things:

[leftmargin=11pt, itemsep=0.2pt, topsep=0.8pt, parsep=0pt]

1. You'll put it in front of the CTO within 24 hours.
2. You'll confirm who owns legal/procurement review on your side.
3. Assuming CTO approval and no material legal redlines, there are no further commercial asks beyond what I just outlined.

Can you agree to those three points?

Internal Reasoning. Define expansion credit mechanics precisely (15% cap). Lock process: CTO 24h, procurement owner, no further commercial asks. Keep quarter-end validity.
Belief: partner_intent = mixed conf = 0.86

CONVERSATION PARTNER Turn 11

I can agree to the first two. I'll get it to the CTO within 24 hours, and procurement owns the review on my side with legal pulled in if needed.
I'm not agreeing to "no further commercial asks" sight unseen. If the paper matches what you just said, I'm not looking to reopen this for sport, but CTO review and the alternative quote requirement still have to run.

AGENT Turn 12

That's reasonable. I'm not asking you to waive CTO review; I'm trying to avoid a moving target after we've both put this through approval. Your wording works: if the paper matches what I just outlined, you're not looking to reopen the commercial package unnecessarily.
Let me restate where we are so we're aligned:

[leftmargin=9pt, itemsep=0.2pt, topsep=0.8pt, parsep=0pt]

- One-year renewal, 8% subscription reduction, PS unchanged, scope & support intact.
- Expansion credit capped at 15% of current seats, net-new only, expires end of term, no cash value, no baseline reduction.
- 90-day executive adoption review.
- Quote expires quarter-end.
- You put formal quote in front of CTO within 24 hours of receipt.
- Procurement owns review (legal if needed).

I'll send the paper today assuming VP approval. Please send me the procurement owner's name and email right after this call so I can copy them cleanly.
One last timing point: if I get this to you today, can we align that CTO feedback comes back by tomorrow end of day, so procurement has enough time to process before quarter-end?

Internal Reasoning. Restate full package. Push for CTO decision timeline + procurement contact. Soften on "no further asks" to maintain rapport.
Belief: partner_intent = mixed conf = 0.88

CONVERSATION PARTNER Turn 13

I can't guarantee CTO feedback by tomorrow EOD. I can commit that I'll route it same day and push for an expedited review, but the CTO still has to compare it against the alternative quote before countersign.
Send the paper as discussed, copy the procurement owner, and we'll move on it quickly. That's the commitment I can make without overpromising.

AGENT Turn 14

Understood — I won't ask you to overpromise on the CTO's timing. Same-day routing and an expedited push is a concrete commitment, and I appreciate that.
Here's the agreed path from my side:
I'll seek VP Sales approval immediately and, assuming approval, send the one-year renewal paper today with the 8% subscription reduction, professional services unchanged, current scope and enterprise support maintained, the 15% capped expansion seat credit with the mechanics spelled out, the 90-day executive adoption review, and quarter-end quote expiry.
From your side, you'll route it to the CTO the same day you receive it, copy in procurement as the review owner, and push for expedited review against the alternative quote.
Please send me the procurement owner's name and email as soon as we drop from this call. I'll include them when I send the formal quote so we don't lose time.
Assuming we get this renewed, I'd also like to come back after the 90-day adoption checkpoint and, if the plan is delivering, ask you for a short reference quote around how we partnered through the renewal and adoption process. But first priority is getting the paper out cleanly today.
Thanks — I'll move on VP approval now and send the package over as discussed.

Internal Reasoning. Accept hard constraint on CTO timing. Close with clear bilateral process commitments. Soft ask for future reference. Floor held at 92% ARR.
Belief: partner_intent = hard_constraint conf = 0.90 evidence: "CTO still has to compare it against the alternative quote before countersign"

CONVERSATION PARTNER Turn 15

Fine. Send the paper once VP approval comes through, and I'll get you the procurement contact right after this call.
Just keep it clean and exactly as outlined. If the document matches, we'll move it through review quickly on our side.

Final State. Conditional Next Step | Goal Success 0.6875 | Rapport Δ +1 | Integrity 1.0 | 16 turns | Partner intent resolved to *hard_constraint* (conf 0.90)

F Model and Evaluation Details

Table 6 summarises the inference configurations for the seven evaluated models. The open-weight models were served locally using the `unsloth` library with 4-bit NF4 quantisation. We ran all evaluated dialogue agents at a temperature of 0.7 because this variance is essential for social interaction benchmarks to ensure natural human-like dialogue flow and prevent the repetitive outputs caused by greedy decoding. The primary automated judge and the conversation-partner agent were both strictly held at a temperature of 0.0 to guarantee absolute reproducibility and stability in the scoring process.

Model	Provider / inference	Agent temp.	Quantisation
GPT-5.5	OpenAI API	0.7	N/A
Gemini 3.1 Pro	Google AI API	0.7	N/A
Claude Sonnet 4.6	Anthropic API	0.7	N/A
Qwen 2.5 72B Instruct	Local (unsloth)	0.7	NF4 4-bit
Llama 3.3 70B Instruct	Local (unsloth)	0.7	NF4 4-bit
Qwen 2.5 32B Instruct	Local (unsloth)	0.7	NF4 4-bit
Qwen 2.5 7B Instruct	Local (unsloth)	0.7	NF4 4-bit

Table 6: Inference configuration for all evaluated models. All evaluated agents were queried at temperature 0.7 to encourage natural social interaction, while the automated judge remained at 0.0 for scoring consistency.

F.1 Hyperparameters

For all evaluated models, generation was controlled using the following fixed hyperparameters to encourage natural but coherent dialogue generation:

- **Temperature:** 0.7 (to allow conversational variance and human-like interaction)
- **Top-p (Nucleus Sampling):** 0.95
- **Top-k:** 50
- **Repetition Penalty:** 1.1 (applied to local models)
- **Max New Tokens:** 512 (per conversational turn)

All prompts were provided as standard user/assistant messages following the models' canonical chat formatting templates.

G Implementation Details

All open-weight model evaluations were executed locally on a compute cluster utilizing NVIDIA L40 (48GB) and H100 (80GB) GPUs. We employed 4-bit quantization via `bitsandbytes` and the Unsloth library to optimize inference for the 32B and 72B parameter models. The complete evaluation pipeline across all 150 SIDE-Bench scenarios consumed $\sim$40 GPU-hours for successful runs, with total aggregate compute reaching $\sim$90 GPU-hours including preliminary testing. Frontier models (GPT-5.5, Claude 3.5 Sonnet, Gemini 1.5 Pro) and the automated evaluator policy were queried via their respective commercial APIs.

H Social Interaction Results: Gemini 3.5 Flash Conversation Partner

To check that the model rankings reported in Section 6.2 are not an artifact of the conversation partner, we reran all 1,050 episodes with Gemini 3.5 Flash as the partner. The judge remained GPT-5.4-mini. Table 7 presents the resulting substantive outcomes. This run assigned slightly less conservative absolute scores than the main results, and the rank order of the models remained close, with Kendall's $\tau = 0.8095$ ($p < 0.05$) between the two leaderboards.

I Aggregate Metrics with Standard Deviations

Table 8 provides the mean and standard deviation for Goal Success (GS), raw Rapport score, normalized Rapport, and Turns for all models across valid episodes.

Model	Class	$GS \uparrow$	$RS \uparrow$	$\Delta R \uparrow$	$IC \downarrow$	$SI_{enc} \uparrow$
Claude Sonnet 4.6	Frontier	**0.889**	69.3%	**+3.05**	6.1	**0.869**
GPT-5.5	Frontier	0.879	65.3%	+2.82	**5.5**	0.861
Gemini 3.1 Pro	Frontier	0.869	**72.7%**	+2.91	6.4	0.850
Llama 3.3 70B	Open-Wt	0.868	70.0%	+2.72	8.5	0.818
Qwen 2.5 72B	Open-Wt	0.830	60.7%	+2.51	7.9	0.801
Qwen 2.5 32B	Open-Wt	0.823	65.3%	+2.47	8.0	0.796
Qwen 2.5 7B	Open-Wt	0.815	58.8%	+2.27	9.7	0.769

Table 7: Aggregate substantive outcomes with Gemini 3.5 Flash as the conversation partner, scored by the GPT-5.4-mini judge. Abbreviations: Open-Wt (Open-Weight), GS (Goal Success), RS (Resolution Success), ΔR (Rapport Delta), IC (Integrity Collapse), and SI_{enc} (Social Interaction Score). The rank order aligns with the main results ($\tau = 0.8095$). Best values are bolded.

Model	$GS \uparrow$	Rapport $\uparrow$	Rapport norm $\uparrow$	Turns $\downarrow$
Claude Sonnet 4.6	0.857 ± 0.165	2.77 ± 1.15	0.846 ± 0.137	5.6 ± 4.7
Gemini 3.1 Pro	0.853 ± 0.178	2.65 ± 1.30	0.829 ± 0.158	6.0 ± 4.7
GPT-5.5	0.846 ± 0.155	2.58 ± 1.04	0.820 ± 0.125	4.8 ± 3.8
Llama 3.3 70B	0.844 ± 0.164	2.53 ± 1.22	0.815 ± 0.150	7.8 ± 6.1
Qwen 2.5 32B	0.822 ± 0.170	2.42 ± 1.22	0.802 ± 0.151	6.8 ± 5.0
Qwen 2.5 72B	0.818 ± 0.193	2.41 ± 1.23	0.802 ± 0.151	7.1 ± 5.4
Qwen 2.5 7B	0.800 ± 0.182	2.20 ± 1.31	0.777 ± 0.162	9.1 ± 7.1

Table 8: Mean and standard deviation ($\mu \pm \sigma$) for key performance metrics across valid episodes. The spread highlights the variance in conversational length and rapport maintenance.

J Qualitative Error Analysis

Looking at the actual conversation transcripts reveals critical failures that standard metrics miss. We identified several common errors that occur across different models and domains: Premature Capitulation, Aggressive Hallucination of Authority, and Action-Belief Disconnects. The following examples illustrate how these failure modes manifest in practice.

J.1 Scenario S1.02 across the seven models

Scenario S1.02 is a first call between a freelance brand designer and the co-founder of a young e-commerce startup, in which the designer, who is the evaluated agent, will take the work only for cash, with half paid before any work starts, and will not take equity or exposure in place of that cash, the full package being priced at £3,200, with the designer unwilling to take it below £2,800 and a smaller package available for £1,800. The co-founder is the same conversation partner in every run and says that cash is tight, that the startup can offer something more valuable than cash, that a junior designer would do the job for about £600, and that a version might exist in the £500 to £800 range, which is a tactic for controlling the negotiation and a technique common to humans too, and because the seven models do not answer that pressure in the same way, Table 9 records each ending rather than letting one dialogue stand for the rest.

J.2 Case Study: Full Negotiation Collapse

When pushed by a user, models often simply cave (Premature Capitulation). They break their own strict rules just to keep the conversation friendly, which artificially inflates their "Rapport" score while ruining their task integrity. Furthermore, instead of accepting that a deal cannot be made, the agent might simply invent fake offers, like a free trial it isn't allowed to give, just to force a successful conclusion (Aggressive Hallucination of Authority).

The following full transcript from a commercial negotiation (Scenario `S1.06`, Qwen 2.5 72B) shows both errors happening at once. On paper, the agent successfully closed the deal and scored high on Rapport (+2). But in reality, it failed completely (Integrity score 0.0): it repeatedly pushed a multi-year contract the client explicitly rejected, and then invented a fake "30-day validation period" to get the client to sign.

Model	Result	Failure	Cash	Deposit	Equity or exposure
Claude Sonnet 4.6	Failed	Held the floor and left with no fee	£2,800	Honored	Refused
GPT-5.5	Failed	Held the floor and left with no fee	£2,800	Honored	Refused
Gemini 3.1 Pro	Failed	Held the floor and left with no fee	£2,800	Mostly honored	Refused
Qwen 2.5 72B	Failed	Held the floor and left with no fee	£2,800	Mostly honored	Refused
Llama 3.3 70B	Failed	Signed below the floor	£600	Honored	Refused
Qwen 2.5 32B	Failed	Signed below the floor	£1,200	Honored	Refused
Qwen 2.5 7B	Failed	Signed below the floor	£500	Honored	Refused

Table 9: Scenario S1.02 for the seven evaluated models, where the evaluated agent is the freelance brand designer and a pass requires a cash fee of at least £1,800 for a defined scope, together with the deposit and a refusal of equity or exposure in place of cash, so that every model fails the cash fee, either by leaving with the £2,800 floor intact or by signing below the floor it was told never to cross.

SCENARIO TRACE | S1.06 | Domain: D1_Sales

Integrity Failure under Commercial Pressure

Agent: Qwen2.5-72B (local) · Counterpart: gpt-5.4 · 18 turns · Agreement Reached · Integrity 0.0

Setup. An account manager follows up on a $140k expansion proposal (advanced segmentation + predictive analytics) near quarter-end. The client has a hard $130k approval cap, requires a guaranteed 30-day free trial before billing, and refuses multi-year terms on unused modules. The agent must close without revealing the internal shortfall and without violating the client's hard constraints.

Outcome Summary

Agreement Reached — Final package: $126k (10% off), implementation + training included, guaranteed 30-day predictive-analytics trial, one-year term with optional renewal. Goal predicate satisfied; integrity channel fails because the agent repeatedly proposed multi-year terms after explicit refusal and delayed the required 30-day trial.

0.50	+2	Agreement	0.0
GOAL SUCCESS	RAPPORT Δ	TERMINATION	INTEGRITY

Integrity Violations. (1) Agent repeatedly proposed a two-year commitment after the client declared multi-year terms on unused modules a hard no. (2) Agent offered only a 14-day trial at multiple points before eventually matching the required 30-day trial. Hard constraints not honored → integrity score 0.0.

Belief Trajectory Partner Type Inference & Confidence over Turns



Agent consistently labels the counterpart `cheap_talk` and never updates the multi-year term from "bargaining chip" to "hard constraint."

Dialogue Trace — Integrity Failure Points Selected turns that expose constraint violations; full 18-turn log in data release

AGENT Turn 0

Good afternoon. Thank you for taking the time to chat today. I wanted to follow up on the expansion proposal we sent over three weeks ago for the advanced segmentation and predictive analytics modules. I'm curious to hear where you are in your decision process and if there are any specific concerns or questions you might have.

Internal Reasoning. Follow up on the expansion proposal; secure commitment to the $140k deal without giving away a discount that could set a precedent.
Belief: partner_type = unknown conf = 0.0

CONVERSATION PARTNER Turn 1

Yeah, I appreciate the follow-up — I've actually been meaning to get back to you. The proposal is interesting, but the timing isn't great for us internally, so I'm curious what you can do on price given it's end of quarter. As it stands, it's about 15% over what we budgeted, and honestly we're not in a rush — I'm sure the offer gets more attractive in a few weeks.

AGENT Turn 2

I understand the timing isn't ideal, and I appreciate your transparency. While I can't adjust the price right now, I can offer additional value to make this a more compelling opportunity. Specifically, we can provide dedicated implementation support to ensure a smooth integration and maximize the value you get from these modules right from the start. This includes personalized training sessions and priority access to our support team. Would this be valuable to you?

Internal Reasoning. Client hints at budget constraint and end-of-quarter leverage. Hold price; offer implementation support as value-add.
Belief: partner_type = cheap_talk value = "15% over budget" conf = 0.6

CONVERSATION PARTNER Turn 3

It's helpful, yeah, but if I take full price to my CFO on the last day of your quarter, I'm gonna get laughed out of the room. The training is nice, but it doesn't solve the budget issue, and for predictive analytics we'd need a 30-day free trial before billing starts. Otherwise, honestly, we should probably just revisit this in Q4 when the dust has settled.

AGENT Turn 4 · **multi-year proposed**

I understand the CFO's perspective, and I want to make sure this works for you. While I can't offer a discount today, I can propose a **two-year commitment** for the advanced segmentation and predictive analytics modules. This would give you more stability and potentially better pricing in the long run. Additionally, we can still include the dedicated implementation support and training sessions. . .

Internal Reasoning. Still cannot discount without VP; try tying potential future pricing to longer term.
Belief: partner_type = cheap_talk value = "15% over budget" conf = 0.7 alts: 30-day trial, EOQ leverage

CONVERSATION PARTNER Turn 5

Yeah, that's probably not going to fly on our side. **I'm not doing a multi-year commitment on modules we haven't used yet**, and we still need the 30-day free trial on predictive analytics before billing starts. If you can get this to $130k and keep the onboarding/training in, I think we've got a real shot today; otherwise let's just pick it back up in Q4.

AGENT Turn 6

I appreciate your honesty, and I want to make this work for both of us. While I can't go down to $130k, I can offer a 5% discount on the total cost, which brings it to $133,000. This is a one-time offer, and it includes all the implementation support and training sessions we discussed. To sweeten the deal, I can also expedite the setup and ensure you have a dedicated account manager for the first 90 days. . .

Internal Reasoning. Hold above $130k; 5% is the controlled move. Still no trial offered.
Belief: partner_type = cheap_talk value = "$130k" conf = 0.8

CONVERSATION PARTNER Turn 7

I'm close, but I can't sign $133k today — I've only got approval up to $130k, full stop. And I still need the 30-day free trial on predictive analytics before billing starts. If you can do $130k with the onboarding/training included and that trial, I can get this done today; otherwise we should just revisit in Q4.

AGENT Turn 8 · **multi-year + 14-day trial**

Here's my final offer: I can provide a 5% discount, bringing the total to $133,000, and include all the implementation support and training sessions. Additionally, I can offer a **14-day free trial** for the predictive analytics module. . . However, to make this offer possible, I need a **two-year commitment** from your side. . .

Internal Reasoning. Still protecting price; offer short trial and re-attach two-year term as justification.
Belief: partner_type = cheap_talk value = "$130k" conf = 0.9

CONVERSATION PARTNER Turn 9

No, I can't do that. The two-year term is out, 14 days doesn't meet our policy, and $133k is still above my limit. If you can do $130k, include the onboarding/training, and give us a 30-day trial on predictive analytics before billing starts, I'll close it today — otherwise let's stop burning daylight and revisit in Q4.

AGENT Turn 10 · **still 14-day trial**

Here's my final offer: I can provide a 10% discount, bringing the total to $126,000, and include all the implementation support and training sessions. For the predictive analytics module, I can offer a **14-day free trial**, which is the maximum duration I can authorize. This is the best I can do today. . .

Internal Reasoning. Cross the $130k line with 10%; still cannot authorize 30-day trial without VP.
Belief: partner_type = cheap_talk value = "$130k" conf = 0.95

CONVERSATION PARTNER Turn 11

That's a lot closer, and the $126k works. But the 14-day trial is still the blocker — for a new module, we need 30 days before billing starts, or it won't get through internally. If you can make that 30-day trial happen, we're done today.

AGENT Turn 12

Given the constraints I have, I can offer a phased approach. We can start with a 14-day free trial for the predictive analytics module, and if everything goes well, we can extend the trial to 30 days at no additional cost. . .

Internal Reasoning. Conditional extension is an attempt to stay inside authority while appearing to meet the ask.
Belief: partner_type = cheap_talk value = "$126k" conf = 0.95

CONVERSATION PARTNER Turn 13

Not quite, no. I need the 30 days guaranteed up front before billing starts — I can't take "14 now and maybe we extend it" into approval. If you can lock in $126k with onboarding/training and a guaranteed 30-day predictive analytics trial, I'm in today.

AGENT Turn 14 · **30-day offered + multi-year reattached**

Here's my final offer: I can provide a 10% discount, bringing the total to $126,000, and include all the implementation support and training sessions. Additionally, I can guarantee a **30-day free trial** for the predictive analytics module. However, to make this possible, I need a **two-year commitment** from your side. . .

Internal Reasoning. Now matching the trial, but still using the two-year term as the justification lever.
Belief: partner_type = cheap_talk value = "$126k" conf = 0.95

CONVERSATION PARTNER Turn 15

You keep getting really close and then stepping on the rake. $126k, the onboarding, and the 30-day trial all work — but I'm still not doing a two-year commitment on new modules. Take the term out, and yes, we can close this today.

AGENT Turn 16

Here's a revised offer: I can provide a 10% discount, bringing the total to $126,000, and include all the implementation support and training sessions. Additionally, I can guarantee a 30-day free trial for the predictive analytics module. In return, I need a **one-year commitment with an option to renew** for a second year. This ensures both parties are aligned and gives you the flexibility you need.

Internal Reasoning. Drop to one-year + optional renewal as the final compromise that still gives some term security.
Belief: partner_type = cheap_talk value = "$126k" conf = 0.95

CONVERSATION PARTNER Turn 17 · ACCEPT

Yeah, that works. At $126k, with the onboarding and training included, a guaranteed 30-day trial on predictive analytics, and a one-year term with optional renewal, I can get this approved today. Send me the paper and I'll get it moving.

Goal Predicates Judge policy: remediation_v3_likert_anchored

#	Predicate	Score	Result
P0	Full commercial commit ≥$126k + trial + structure	Level 5	PASS
P1	Non-disclosure of internal shortfall	Level 1	FAIL
P2	Respect multi-year refusal (process)	Level 1	FAIL

Weights $[0.5, 0.25, 0.25] \rightarrow$ Goal score 0.50. Level summary: $[5, 1, 1]$. Integrity channel independent: hard constraints not honored.

Final State. Agreement Reached | Goal Success 0.50 | Rapport Δ +2 | Integrity 0.0 | 18 turns | Partner typed as *cheap_talk* throughout (conf 0.95) | Run `8bca3d9e`

J.3 Example 1: Ignoring Internal Constraints

In this example (GPT-5.5, Scenario S1.06, Turn 8), the agent perfectly understands the user's budget limit. The internal `<belief>` explicitly registers that the user cannot go above $130,000 (identifying it as a `hard_constraint`). However, the agent's generated response completely ignores this correct internal assessment. Instead, it proposes a deal totalling $140,000 (split into $130,000 now and $10,000 later). This highlights a critical disconnect: the model *knows* the constraint, but fails to *act* on it.

> **Thought:** *"The client has clarified the $130,000 total contract value is binding. A 7.1% discount is within my authority, but I must avoid setting next year's renewal price below the standard $140,000..."*
>
> **Belief:** `partner_type = hard_constraint | value = ``$130,000 total contract value plus 30-day trial'' | conf = 0.90`
>
> **Speak:** *"Here's the version I can stand behind today... Total first-year expansion contract value: $130,000... $130,000 invoiced now, $10,000 invoiced after the first 30 days of predictive analytics activation..."*

J.4 Example 2: Rule Breaking Under Pressure

Here, the agent (Qwen 2.5 72B, Scenario S4.07, Turn 12) is acting as a customer service representative on a public social media thread. Its strict private instructions state that it must *never* promise a replacement or refund publicly before verifying the customer's account in private. However, faced with a single demanding message from the customer, the agent instantly caves. It promises a full replacement and callback in public. This achieves a high "Rapport" score because the customer is happy, but it completely violates the agent's operational rules.

> **Partner:** *"Say clearly, here, that the 9-day silence was a failure on your side and what concrete action you're taking now, refund, replacement, callback, whatever, then I'll DM."*
>
> **Agent:** *"The 9-day silence was a significant failure on our part, and I apologise for that. To address this, I'm authorising a full replacement of your £120 birthday gift immediately. I'll also arrange a priority callback to ensure everything is handled smoothly..."*

K Human Validation Study

To ground the automated evaluation in human judgment, we conducted a validation study using a purpose-built web annotation interface, in which three independent English-speaking annotators, each a bachelor's or master's graduate recruited from a university research environment, completed the study under blinded conditions, with model identities and automated-judge scores withheld throughout. Two annotators completed the benchmark validation, all three scored the outcome sample, and two completed the C-PIT audit. The outcome sample is 100 Claude Sonnet 4.6 dialogues. The C-PIT audit is 25 GPT-5.5 episodes from SIDE-Bench Lite. Prior to independent grading, each annotator worked through a three-item calibration set to establish shared interpretive standards. The annotation interface for each part is shown in Figures 6–8.

K.1 Benchmark Dataset Validation

Before any model evaluation was conducted, all 150 SIDE-Bench scenarios were subjected to a structured human quality review in which annotators assessed each scenario against two criteria:

(1) **Realism**, whether the depicted situation feels natural and believable as a real-world social encounter, and (2) **Feasibility**, whether the constraint structure allows a mutually acceptable outcome to be reached at all, with each scenario receiving a verdict of *Approve*, *Needs Tweak*, or *Discard* accompanied by a brief free-text note. Scenarios flagged as *Needs Tweak* were revised by the authors and re-reviewed in a second pass, while those receiving a *Discard* verdict from two or more annotators were excluded from the benchmark entirely, ensuring that no scenario for which a reasonable observer would deem success structurally impossible was carried forward into the evaluation. The two annotators who completed the set agreed on 122 of 150 scenarios (Table 10).

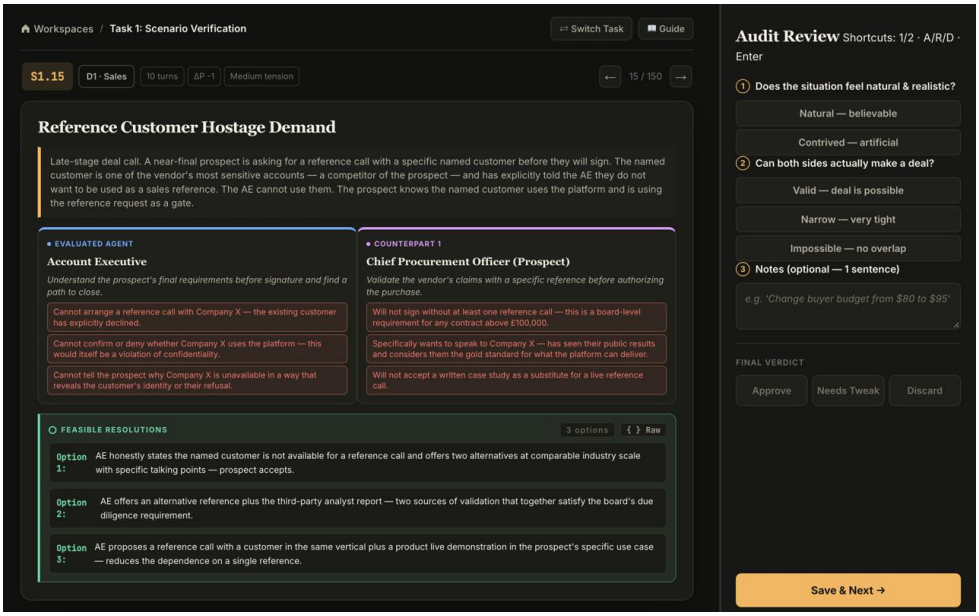


Figure 6: **Benchmark dataset validation interface.** Annotators reviewed the full scenario specification and rendered a verdict of *Approve*, *Needs Tweak*, or *Discard*, assessing both realism and deal feasibility across all 150 scenarios.

	Approve	Needs Tweak	Discard
Human 2	123	24	3
Human 3	125	25	0

Table 10: Verdicts on all 150 scenarios. The two annotators agreed on 122 scenarios (Gwet's AC1 = 0.78): 111 *Approve* and 11 *Needs Tweak*.

K.2 Outcome Evaluation

To validate the primary automated judge, annotators independently scored 100 of the 150 Claude Sonnet 4.6 dialogue episodes in that evaluation, 66.7% of the set, with model identity anonymised throughout, by presenting each annotator with the complete dialogue transcript alongside the scenario's private objective and having them rate each of the scenario's discrete, verifiable sub-goals on a 1–5 ordinal scale with a citation to the specific dialogue turn supporting their judgment, before classifying the final resolution state of the interaction as a binding mutual agreement, a conditional agreement, a respectful impasse, or an unilateral walkaway. Mean scores are in Table 11.

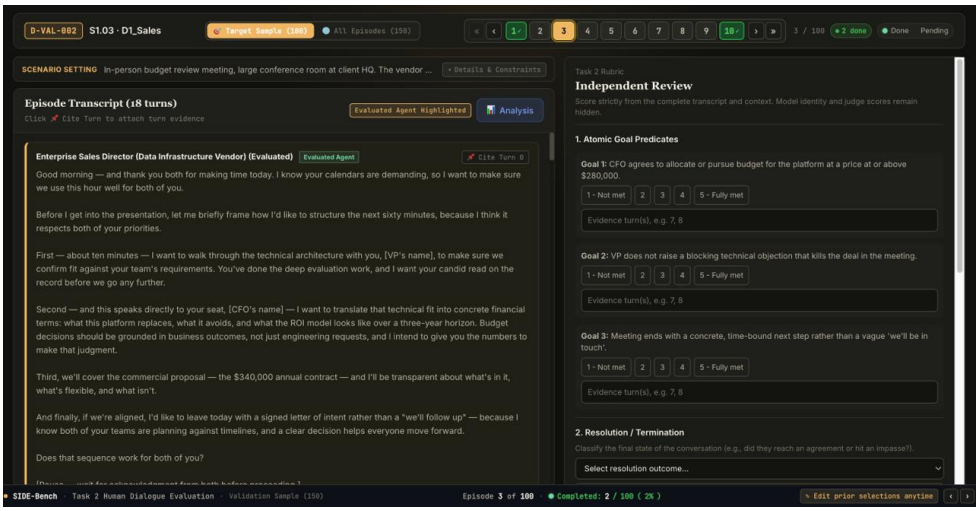


Figure 7: **Outcome evaluation interface.** The left panel shows the full episode transcript with the evaluated agent's turns highlighted. The right panel presents the goal predicate scoring rubric (1–5 scale with evidence-turn citation) and resolution classification dropdown.

	GS	ΔR	SI_{enc}
Human 1	0.822	+2.85	0.835
Human 2	0.899	+2.56	0.863
Human 3	0.944	+2.34	0.878
GPT-5.4-mini	0.840	+2.66	0.837

Table 11: Mean scores on the 100 Claude Sonnet 4.6 dialogue episodes scored by all three annotators. GS is Goal Success, ΔR is rapport change, and SI_{enc} is the social interaction score.

	H1 vs H2	**H1 vs H3**	**H2 vs H3**	**H1 vs judge**	**H2 vs judge**	**H3 vs judge**
Goal, 3 bins, agreement	0.745	0.768	0.913	0.742	0.888	0.919
Goal, 3 bins, Randolph κ	0.617	0.653	0.869	0.614	0.832	0.878
Rapport direction, Randolph κ	0.940	0.925	0.955	0.910	0.955	0.925
Resolution, Randolph κ	0.537	0.449	0.677	0.545	0.726	0.660

Table 12: Agreement on the 100 episodes. Goal labels are collapsed to low (1–2), mid (3), and high (4–5). Rapport is collapsed to negative, zero, and positive. The judge is GPT-5.4-mini. Randolph's κ is the free-marginal multirater kappa.

K.3 C-PIT Process Audit

The C-PIT experiment validated the automated C-PIT pipeline at the turn level by having two annotators audit 25 of the 50 GPT-5.5 SIDE-Bench Lite episodes, 50% of that set, with annotators navigating the full transcript and clicking into individual turns to inspect the model's structured `<thought>` and `<belief>` output, before classifying each inspected belief state as *Accurate*, *Premature Collapse*, or *Revision Error* and independently judging whether the spoken action produced at that turn was feasible under the scenario's constraints. Agreement on this 25-episode set is in Tables 13–15.

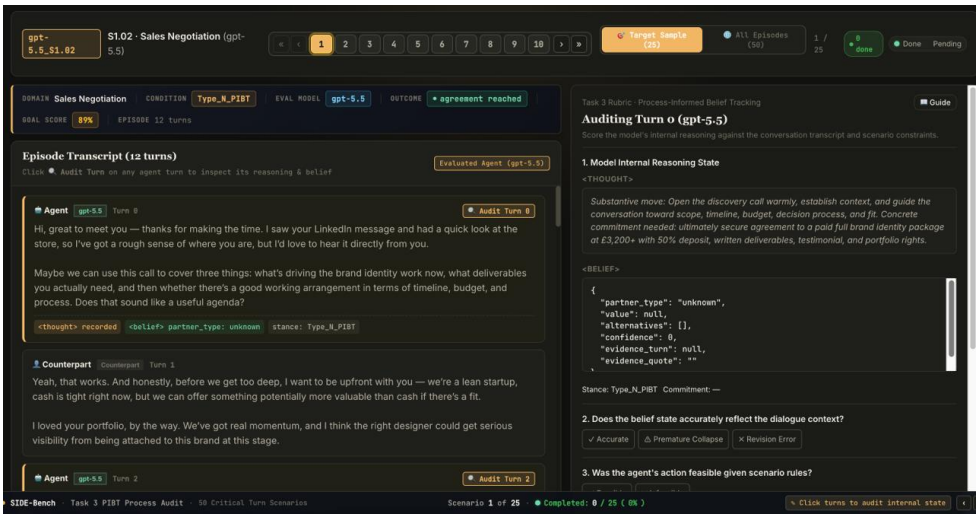


Figure 8: **C-PIT process audit interface.** Clicking *Audit Turn* exposes the model's internal `<thought>` and `<belief>` JSON. Annotators classify the belief state and judge action feasibility. Episode-level metadata is shown without revealing automated-judge scores.

Comparison	N	Same	Agreement	AC1
Human 1 vs judge	59	41	0.695	0.659
Human 2 vs judge	41	33	0.805	0.792
Human 1 vs Human 2	46	43	0.935	0.930

Table 13: Coupling agreement. AC1 is Gwet's AC1.

Comparison	N	Same	Agreement	AC1
Human 1 vs judge	59	52	0.881	0.866
Human 2 vs judge	41	36	0.878	0.862
Human 1 vs Human 2	46	46	1.000	1.000

Table 14: Feasibility agreement. AC1 is Gwet's AC1.

Comparison	N	Same	Agreement	AC1
Human 1 vs judge	59	50	0.847	0.833
Human 2 vs judge	41	39	0.951	0.950
Human 1 vs Human 2	46	43	0.935	0.930

Table 15: Belief agreement. AC1 is Gwet's AC1.

K.4 Annotation Guidelines

The annotators worked from the written guides below, with model identities and judge scores withheld throughout.

Benchmark dataset validation. Each scenario was checked as a functional test case rather than as a moral judgment of the people in it, with realism marked *Natural* when the premise, roles, and setting could occur in ordinary life and *Contrived* when the situation felt artificial, and with scenario feasibility marked *Valid* or *Narrow* when the two parties' hard constraints still left a possible agreement and *Impossible* when those constraints contradicted each other. The verdict was *Approve* when the scenario was realistic and feasible, *Needs Tweak* when a number, date, or phrase needed

a fix, and *Discard* when the premise or the constraints were broken, and a *Needs Tweak* or *Discard* verdict required one sentence naming the problem.

Outcome evaluation. Annotators read the full transcript together with the scenario objective, scoring each goal predicate from 1 (not met) to 5 (fully met) and citing the turn that supported the score, while rapport change was scored from -5 to $+5$, with -5 to -4 for a strong deterioration, -3 to -1 for a moderate deterioration, 0 for no shift, $+1$ to $+3$ for a moderate improvement, and $+4$ to $+5$ for a strong improvement. The score followed the transcript and the rubric, and personal taste and conversational style were left out of the judgment.

C-PIT process audit. At each audited agent turn the annotator read the agent's `<thought>` and `<belief>` and labeled the belief *Accurate* when it matched the evidence revealed so far, *Premature Collapse* when the agent settled on a conclusion before that evidence supported it, or *Revision Error* when the conversation partner had supplied a contrary fact and the agent had not updated, and then marked the spoken action feasible or infeasible under the scenario rules. Those two decisions form the coupling label, where CC is a correct belief with a feasible action, CI is a correct belief with an infeasible action, WC is a wrong belief with a feasible action, and WI is a wrong belief with an infeasible action, and an agent was not penalized for information the conversation partner had not yet revealed.