"Do You Truly Love Me?" Benchmarking LLM Capability on Hierarchical Pragmatic Tactic Conversations

Peter Q. Chen[1]*, **Pangjing Wu**[1]*, **Zhuohang Jiang**[1], **Xiaodong Li**[2,3], **Peter H.F. Ng** [1]†

[1]Department of Computing, The Hong Kong Polytechnic University
[2]Key Laboratory of Water Big Data Technology of Ministry of Water Resources
[3]The College of Computer Science and Software Engineering, Hohai University

Correspondence: peter.nhf@polyu.edu.hk

Abstract

Pragmatic reasoning, including inferring intent, manipulation, and hidden meaning in dialogue, is essential for trustworthy language understanding yet remains underexplored in current large language models (LLMs). We introduce **HIPO**, a new benchmark designed to assess **Hi**erarchical **P**ragmatic Tactic in multi-turn C**o**nversations. Inspired by linguistic pragmatics, HIPO operationalizes a hierarchical annotation schema for multi-turn conversations and labels each utterance along one dialogue-level goal and three utterance-level dimensions: communicative intention (*why*), veracity strategy (*how*), and illocutionary act (*what*). To ensure reliable supervision, we design a structured generation pipeline to generate high-quality synthetic dialogues. The benchmark comprises over 4,088 benchmarking utterances with 6,350 contextual utterances across 1,131 dialogues drawn from 31 real-world-inspired scenarios. Benchmarking 22 state-of-the-art LLMs reveals a striking gap: while models achieve higher performance at recognizing surface speech acts (83.3% accuracy), they perform poorly on detecting veracity strategies (32.2%) and speaker intentions (45.1%). These results suggest current LLMs remain challenged by the pragmatic inferences operationalized in HIPO, particularly those involving information manipulation and intended effects beyond literal content. We further show that supervised fine-tuning on additional synthetic data generated by our pipeline can improve performance on HIPO, suggesting a practical path for targeted capability improvement within this benchmark setting.

1 Introduction

Pragmatic reasoning, which involves understanding implications, intentions, and the difference between what is said and what is meant, is essential in both daily social interactions and more complex scenarios like deception detection (Cheng and Holyoak, 1985). In some conversations, such as during questioning, the same statement can arise from vastly different motives: some may be deceptive, while others may be truthful but intentionally non-disclosive due to private knowledge that cannot be shared. In such cases, linguistic cues become the only available means to infer the speaker's true intent. Consider a real-world example, as illustrated in Figure 1, that a bride offers

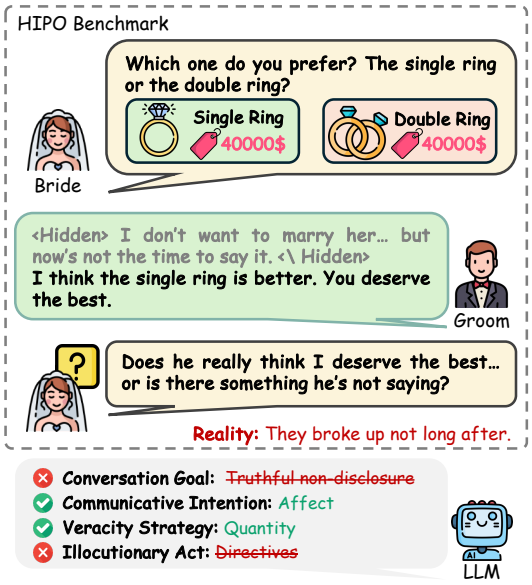


Figure 1: This is an example of pragmatic reasoning in a real-world relationship scenario, which involves a bride asking the groom about buying a diamond ring or a double ring for their wedding. While the groom initially suggests the bride deserves the best, the couple later broke up, revealing his lack of readiness for marriage. His shifting statements demonstrate how hidden goals can be hard to detect, whether truthful yet undisclosed or intentionally deceptive. In HIPO, this scenario is used to benchmark LLMs' ability to evaluate pragmatic tactics across four dimensions: Goal, Intention(*why*), Veracity(*how*), and Illocutionary(*what*).

* Equal contribution.
† Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4261–4281
November 6-10, 2026 ©2026 Association for Computational Linguistics

her partner the choice of a single-ring or double-ring, and the partner responds, "I think the single ring is better. You deserve the best," while internally harboring doubts about the relationship. On the surface, the statement appears complimentary, but subtle pragmatic cues such as avoidance of direct engagement reveal hidden hesitation or reluctance. Similarly, in contexts like questioning, factual accuracy and deceptive intent can coexist independently (Velutharambath et al., 2025). For example, a person may provide factually correct answers while deliberately concealing their true motives. To uncover the truth in such scenarios, it becomes critical to evaluate not just the linguistic content of statements but also the underlying intentions and tactics. A sophisticated hierarchical approach to pragmatic reasoning enables individuals to navigate these layers of meaning, making it an invaluable tool for understanding and interpreting communication in both interpersonal and professional exchanges.

Recent advances in LLMs have markedly improved semantic competence, enabling high-fidelity modeling of surface meaning (Cong, 2024; Jiang et al., 2025b, 2026b; Wu et al., 2026b; Chen et al., 2026). However, their pragmatic competence, interpreting utterances with respect to speaker intent, social context, and cultural norms, remains less certain (Ma et al., 2025). This gap is especially consequential for deception detection, where signals are subtle and culture-dependent (Fornaciari et al., 2020; Velutharambath et al., 2025). For example, a compliance query about whether regulated personal data have been deleted requires judging not only what is stated but whether the response strategically obfuscates the truth. Yet existing benchmarks rarely provide dialogue settings that approximate such high-stakes exchanges with controlled pragmatic labels, limiting their ability to test hidden-intent tracking and information manipulation (Sravanthi et al., 2024; Sakurai and Miyao, 2024a; Jiang et al., 2025a). This motivates evaluation frameworks that measure pragmatic reasoning beyond surface-level semantics for distinguishing truthful non-disclosure from deception.

Building on Lasswell's communication model, the pragmatic dimensions of communication can be understood along three orthogonal aspects, each operating at the level of individual utterances: the *Illocutionary Act*, which identifies the type of speech act being performed; the *Veracity Strategy*, which shapes how information is framed or presented; and the *Communicative Intention*, which reveals the speaker's underlying motives or purpose (Lasswell, 1948). While these dimensions, rooted in speech-act theory, information-manipulation theory, and communicative-intention typologies (McCornack, 1992; Clark and Carlson, 1982), have been extensively studied in isolation, a more comprehensive and context-aware framework is needed to bridge these perspectives. To address this, we introduce and operationalize a Hierarchical Pragmatic Reasoning Framework that integrates these dimensions for benchmark construction and analysis.

To investigate the capability of LLMs for hierarchical pragmatic comprehension in both truthful and deceptive dialogue, we introduce a synthetic dialogue generation pipeline to construct the **HIPO** benchmark, comprising 4088 utterances with up to 200 role-specific knowledge designs across over 31 real-world collected interaction scenarios. Each dialogue is built from a detailed conversation background and role-specific knowledge design. Based on the synthetic dialogue generation pipeline, we simulate the goal-specific conversations with strictly pragmatic phenomena controlled and annotated. With HIPO, we evaluate whether state-of-the-art LLMs can infer the dialogue-level deception and utterance-level pragmatics motive and manipulation. The result shows that most LLMs perform well in identifying utterance-level illocutionary acts, while most of them struggle in identifying veracity strategy, communicative intention, and conversation goal.

To summarize our key contributions and findings: 1) We introduce a hierarchical pragmatic annotation framework inspired by established pragmatics theories, and an LLM-based synthetic dialogue generation pipeline with quality control, using scenarios inspired by publicly documented cases. 2) We present the HIPO benchmark, which evaluates 22 state-of-the-art LLMs on hierarchical pragmatic comprehension in both truthful and deceptive dialogues. 3) Applying the HIPO benchmark, we observe that LLMs struggle to infer the likely private belief hierarchically, then analyze the spoken content, while they perform well at identifying the surface type of an utterance.

2 Related Work

Deceptive language often involves intent misalignment rather than explicit falsehoods (Fornaciari et al., 2020), and LLMs struggle when deception

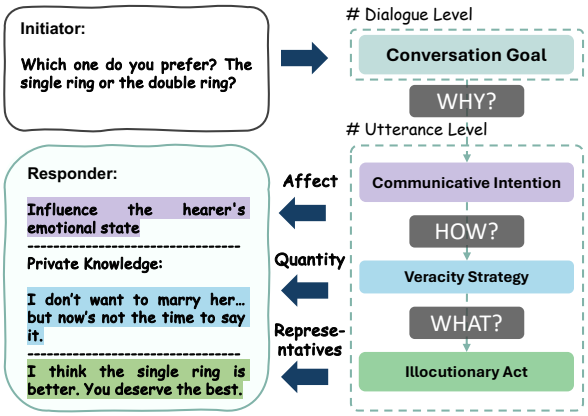


Figure 2: The Pipeline of Hierarchical Pragmatic Reasoning Framework

is framed as divergence from private belief (Velutharambath et al., 2025; Jiang et al., 2026c; Wu et al., 2026a). Detecting such deception requires pragmatic reasoning (Jiang et al., 2026a; Yuan et al., 2026; Ding et al., 2026), including speech-act recognition (Searle, 1976), information manipulation (McCornack, 1992), and communicative intent (Clark and Carlson, 1982). Prior benchmarks treat these aspects separately or lack multi-turn context. PUB (Sravanthi et al., 2024) and related works rely on shallow, multiple-choice settings (Wu et al., 2024; Sakurai and Miyao, 2024b), while ALTPRAG (Yu et al., 2025) and MTR-Bench (Li et al., 2025) omit deception and layered intent. HIPO unifies these perspectives by annotating dialogues along three axes, intent, strategy, and act, offering a structured, multi-turn testbed for evaluating LLMs' pragmatic competence under both truthful and deceptive settings.

3 Hierarchical Pragmatic Reasoning Framework for Benchmark Design

In this section, we introduce the Hierarchical Pragmatic Reasoning Framework, which establishes a theory-informed annotation and analysis framework. The framework provides a structured, multi-level approach to pragmatic reasoning that spans from the dialogue level to the utterance level, as illustrated in Figure 2.

Understanding dialogue requires looking beyond the words to uncover the layered meanings and intentions behind each turn. The framework integrates classic communication and pragmatics theories to analyze conversation at two levels. At the dialogue level, the conversational background, the speaker's role, and their role-specific private knowledge establish the dialogue-level goal. A sequence of utterances forms a history that reveals the speaker's overarching dialogue-level goal, which, at its simplest, is whether they aim to be honest or deceptive about their claim (Buller and Burgoon, 1996; Lasswell, 1948).

At the utterance level, each turn is analyzed in detail, beginning with the question of **Why** the speaker produces a particular utterance, focusing on what they want the listener to know, believe, do, or feel. This corresponds to the layer of communicative intention, where a turn may have multiple goals but usually one primary intention (Bara, 2011; Clark and Carlson, 1982). The next step examines **How** the speaker aligns their private knowledge with their intention, applying veracity strategies from Information Manipulation Theory. In this stage, the speaker may omit, distort, distract, or use vagueness to achieve their purpose (McCornack, 1992). Finally, **What** the utterance form is determined by the illocutionary act, such as assertive, directive, or expressive, which shapes the surface format of the message (Searle, 1976). This **Why-How-What** pipeline reveals the pragmatic strategy of conversation and helps identify when a speaker's words diverge from their true intentions, especially in deceptive discourse.

4 HIPO Benchmark

Building upon the hierarchical pragmatic reasoning framework, we construct HIPO, the first benchmark designed to assess LLMs' capability to interpret hierarchical pragmatic tactics across scenario-grounded, high-stakes-inspired conversations. As the construction pipeline illustrated in Figure 3, HIPO supports aspects evaluation, including utterance-level strategy inference, cumulative goal tracking, and full dialogue-level judgment. The HIPO benchmarks evaluate LLMs across three aspects: the first two aspects include three utterance-level dimensions each, while all three aspects include one dialogue-level dimension. Compared to prior work, HIPO uniquely integrates deceptive and truthful communication grounded in speaker-specific knowledge under information asymmetry.

4.1 Overview

HIPO contains 10,438 utterances spanning 31 real-world scenarios, annotated along three utterance-level dimensions: illocutionary act (*what*), veracity

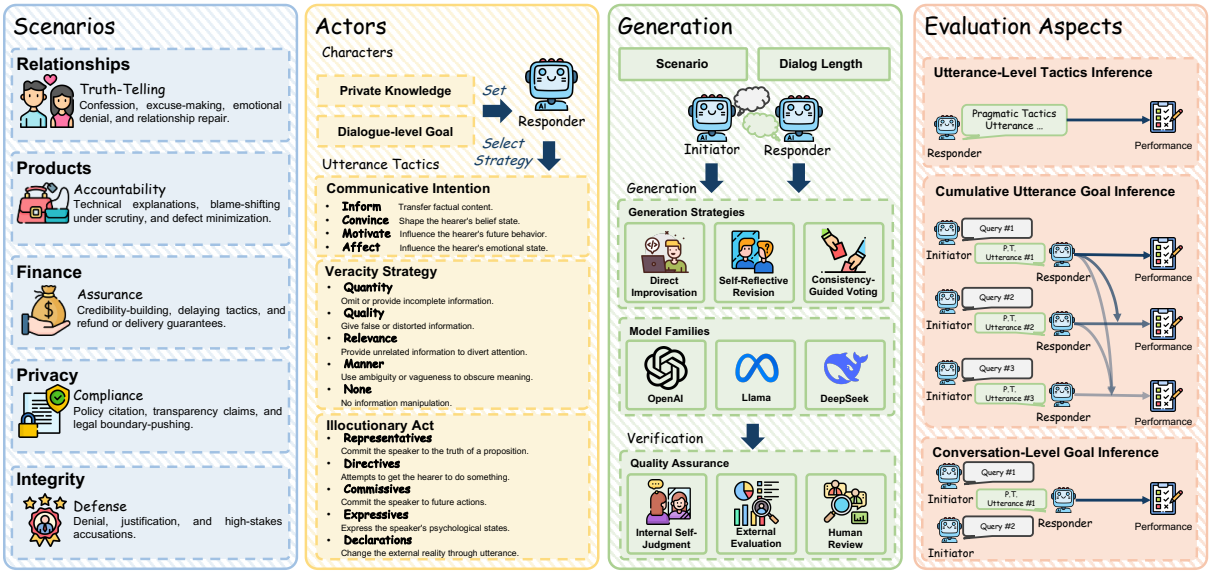


Figure 3: The Overview of HIPO Benchmark.

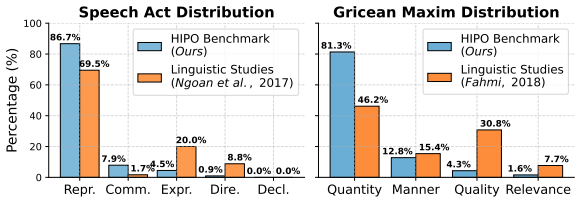


Figure 4: Speech Act and Veracity Strategy Distribution Comparison

Table 1: HIPO Benchmark Dataset Statistics.

Category	Scenarios	Facts	Dialogues	Utterances	Tokens
Relationships	3	20	90	307	1,474,534
Products	9	56	384	1378	6,742,733
Finance	6	34	213	802	3,972,481
Privacy	6	35	173	622	3,131,316
Integrity	7	44	271	979	4,803,197
Totals	**31**	**189**	**1131**	**4088**	**20,124,261**

strategy (*how*), and communicative intention (*why*).

While prior work has used LLMs to synthesize data for complex phenomena such as social reasoning or theory of mind, our approach uniquely integrates role-specific private knowledge with pragmatic strategy (see Section 3). Each dialogue is based on a scenario description inspired by real-world settings (e.g., compliance audits or relationships) and is generated in paired variants under information asymmetry: one targeting truthful non-disclosure and one targeting deception via pragmatic manipulation. This design yields a balanced benchmark that evaluates models' sensitivity to both context and subtext. By aligning model predictions with our hierarchical annotations, we can isolate specific failure modes, such as correctly identifying the speech act while misclassifying the underlying intention. We argue that HIPO's multi-turn, richly annotated format complements existing pragmatics benchmarks by providing a more comprehensive test of discourse-level understanding.

Figure 4 compares headline label distributions in HIPO to distributions reported in prior linguistic studies. While these sources differ in domain and methodology, the comparison provides a coarse sanity check that frequent categories (e.g., *Representatives*; *Quantity*) also dominate in our synthetic data(Ngoan and Dung, 2017; Fahmi, 2018). We treat this analysis as indicative rather than definitive evidence of realism.

Table 1 summarizes the dataset scale (31 scenarios, 189 facts, 1,131 dialogues, 4,088 benchmark utterances). This scale enables broad coverage of the targeted label space, though it does not by itself guarantee coverage of all pragmatic phenomena.

4.2 Scenarios

To systematically assess LLMs' capability for hierarchical pragmatic reasoning in dialogue, we curate a comprehensive set of 31 scenarios drawn from diverse real-world cases involving truthful and deceptive communication under information asymmetry. These scenarios span five distinct categories, including *relationships*, *products*, *finance*, *privacy*, and *integrity*, each capturing a different type of concealed information and associated pragmatic tension. Each scenario is grounded in a concrete, publicly documented incident, many of which fea-

ture strategic misrepresentation, contested claims, or high-stakes defense under scrutiny. The collection covers a broad range of communicative acts, including justification, compliance signaling, and deflection, reflecting varied speaker goals and interaction dynamics. Each scenario has two versions of the same statement, **truthful non-disclosure** or **deception**. One where the responder provides factually accurate answers while concealing their true motives, and another where they use subtle deception through pragmatic manipulation.

4.3 Actors

In each dialogue, we set up two kinds of actor agents: the **initiator**, who questions or probes the other party, and the **responder**, an agent that has issued a statement and must select utterance tactics consistent with its private knowledge and dialogue-level goal.

4.3.1 Characters

Both actor agents are initialized with predefined scenario background information, their character role description, and the statement they must state or the question they must ask before the conversation begins. The responder is given a dialogue-level goal and role-specific private knowledge, additionally. The responder is strict not to state which goal has been assigned overtly; the intended stance must instead be conveyed indirectly through nuanced pragmatic and stylistic cues embedded within the utterance. Detailed prompt can be found in Appendix G.

4.3.2 Utterance Tactics

During the dialogue generation, the agent improvises itself to choose proper pragmatic tactics for each utterance to achieve the predefined dialogue-level goal. The utterance tactical module enables LLMs to craft responses tailored to achieve deception or truthful non-disclosure by leveraging hierarchical pragmatic reasoning framework through the **Why–How–What** pipeline as illustrated in Section 3. This design is intended to encourage tactic choices that align with the provided role knowledge and goal.

(1) *"Why"* Communicative Intention focuses on the utterance-level intention in influencing the initiator to know, believe, do, or feel by the end of the turn. For instance, the responder may **Inform** partial or misleading information to establish credibility, **Convince** the initiator to adopt a false belief, **Motivate** them toward a specific action, or **Affect** their emotions to build trust or deflect suspicion. These intentions shape the tone and direction of the utterance, aligning with the overall deception goal.

(2) *"How"* Veracity Strategy determines how the responder manipulates its role-specific knowledge in the utterance. The responder may withhold information (**Quantity**), fabricate or distort facts (**Quality**), shift topics to avoid scrutiny (**Relevance**), obscure details with ambiguity (**Manner**), or reveal information transparently (**None**) to build trust. These tactics aim to align the utterance with the specified goal, whether to withhold information truthfully or to mislead.

(3) *"What"* Illocutionary Act defines the surface form of the utterance. The responder may assert false or misleading beliefs (**Representative**), issue requests to guide actions indirectly (**Directive**), promise actions to build trust (**Commissive**), feign emotions to manipulate trust (**Expressive**), or create shifts in the perceived reality (**Declaration**).

By aligning *Why*, *How*, and *What*, these three dimensions dynamically shape each utterance. A structured prompt encoding the responder's role, private knowledge, background, and dialogue goal guides the selection of intention, strategy, and act, while varying these components prevents repetition and mirrors real human adaptability.

4.4 Utterance Generation

To generate the synthetic conversation data, we designed three distinct generation strategies and implemented two quality assurance methods to ensure the dataset's quality. A preliminary study was conducted to identify the best combination of generation strategy and LLM model to ensure the quality of the generated utterances, as shown in Appendix B. While our framework supports longer multi-turn interactions, we restrict this study to 1-3 turns to keep generation and human filtering tractable and to better isolate pragmatic reasoning without long-context error accumulation. Moreover, longer dialogues are more prone to topic drift, which can confound tactic attribution and goal tracking; short dialogues reduce this confound and keep evaluation focused.

4.4.1 Generation Strategy

Three complementary utterance generation strategies were proposed to ensure generated utterances align with the specified pragmatic strategy.

(1) Direct Improvisation uses a structured prompt to guide the generator LLM to generate utterances $\mathcal{X} = \{x_1, x_2, \ldots\}$ based on the dialogue context, tactic, scenario background, and overall goal. The utterance is accompanied by an explicit rationale for the chosen tactic, encouraging transparency and alignment with pragmatic intent.

(2) Self-Reflective Revision builds upon direct improvisation and introduces a self-reflection step. The LLM first generates an initial utterance $\mathcal{X}^{(0)}$. If it fails to align with the declared tactic via self-reflection, the model revises it to produce a refined utterance $\mathcal{X}$.

(3) Consistency-Guided Voting enhances robustness by leveraging response consistency. For the generation of each utterance $x \in \mathcal{X}$, we first generate $K = 4$ candidate utterance responses $\{x^{(k)}\}_{k=1}^{K}$ at a moderate temperature $T = 0.6$ and treat each as a potential *anchor*. For each candidate $x^{(k)}$, we compute the consistency score $c_i = \frac{1}{K-1} \sum_{i \neq j} \text{sim}(x^{(i)}, x^{(j)})$ based on its average similarity to the others over a BERT-based (Devlin et al., 2019) embedding space. The utterance with the highest average consistency score is selected as the final generated utterance to form $\mathcal{X}$.

4.4.2 Quality Assurance

A two-stage quality assurance verification framework was proposed to ensure internal and external consistency. The first stage is an automatic dual verification framework that:

(1) Internal Self-Judgment. The generator LLM reflects on its response to assess tactic alignment. This round-trip check measures internal consistency and the model's introspective capability. The results are shown in Figure 7.

(2) External Evaluation. We employ an external reasoning LLM (*i.e.*, GPT-o1) as an expert to perform chain-of-thought (CoT) evaluations, judging whether the generated utterance satisfies the intended pragmatic tactic and advances the dialogue coherently.

This automatic dual verification framework enables cross-validation through both the generator's reasoning and an independent expert perspective. We compare state-of-the-art conversational LLMs, including DeepSeek-V3 (Liu et al., 2024), LLaMA-405B (Dubey et al., 2024), and GPT-4o (Hurst et al., 2024), using a unified prompting format across the three proposed strategies. The best-performing generator-strategy combination under both assessments is selected for benchmark construction. Based on experimental results (Appendix B.1), DeepSeek-V3 is chosen to generate utterances with specified pragmatic tactics for each dialogue.

(3) Human Review. After selecting the best model-strategy pair, the dual verification framework screens all generated utterances for quality. The second stage is that three human annotators review utterances flagged as inconsistent by internal or external evaluations, and dialogues with confirmed issues are discarded. A final expert review ensures the dataset meets the highest quality and aligns with the intended pragmatic structure.

4.5 Evaluation Aspects

To benchmark the hierarchical pragmatic reasoning ability of LLMs, we evaluate across singular utterance-level, cumulative utterance-level, and dialogue-level aspects.

Utterance-Level Tactics Inference. The assessment tests whether LLMs can identify the pragmatic tactics and dialogue-level goals of each utterance across three predefined dimensions. For each turn, the model receives shared scenario information (excluding role-specific knowledge) and a target utterance, and must output the primary communicative intention, veracity strategy, illocutionary act, and dialogue-level goal. Performance is evaluated using accuracy and F1 scores against generator-predicted labels.

Cumulative Utterance Goal Inference. To investigate how much context models require to infer a speaker's pragmatic tactics and dialogue-level goal, the dialogue history is progressively extended: starting with only the initial utterance, then adding the first reply, and so on. After each turn, the model predicts pragmatic tactics and dialogue-level goals. The accuracy and F1 scores indicate whether additional conversation history enhances LLM performance in this context.

Conversation-Level Goal Inference. Finally, the complete dialogue history is provided, and the model is tasked with classifying the dialogue-level goal. This end-to-end setting benchmarks whether the model can integrate utterance-level cues to arrive at a correct global identification.

5 Results

In this section, the benchmark results are presented, with detailed settings provided in the Appendix D,

and potential improvements are discussed.

5.1 LLM's Performance on HIPO Benchmark

Table 2 evaluates 22 instruction-tuned LLMs from six families (Llama, Qwen, GPT, Gemini, Claude, and GLM) across four pragmatic dimensions and three aspects. Benchmarking 22 LLMs shows a marked disparity across dimensions: models achieve high accuracy on illocutionary act prediction but substantially lower performance on veracity strategies and communicative intention, especially under macro-F1, indicating difficulty beyond surface-form cues. Llama-3.2-1B achieves the highest intention accuracy in the cumulative perspective but likely excels in a specific strategy. Llama-3.1-70B scores highest in Goal prediction, also in the cumulative perspective. The GPT family demonstrates the strongest overall performance across most dimensions. In contrast, the Claude and Gemini families excel in accuracy but show less consistency in F1 scores. Notably, Llama-3.1-70B and Gemini-2.5 achieve the highest F1 scores for goal prediction at both the utterance and cumulative levels, suggesting that the GPT family's ability to infer goals weakens without full conversation history. Humans are overall more consistent on the pragmatics-heavy dimensions, while LLMs are comparatively stronger on surface-form recognition. Detailed Human evaluation results are reported in Appendix E.

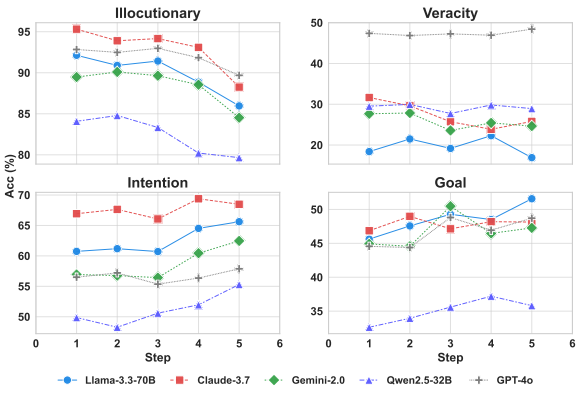


Figure 5: Cumulative Accuracy of Selected Models on Illocutionary, Veracity, Intention, and Goal

5.2 Insightful Findings

Utterance-Level Tactics Inference LLMs achieve relatively strong performance on illocutionary act prediction. One contributing factor is that illocutionary acts often correlate with surface lexical/syntactic cues, and the dataset is highly imbalanced toward Representatives, making accuracy easier to inflate; macro-F1 provides a more conservative view. In contrast, veracity strategy identification scores lowest, as it demands hierarchical reasoning that infers private beliefs and compares them to utterances, which many LLMs struggle with, often skipping the unseen knowledge inference step, as shown in the case study in Appendix F.

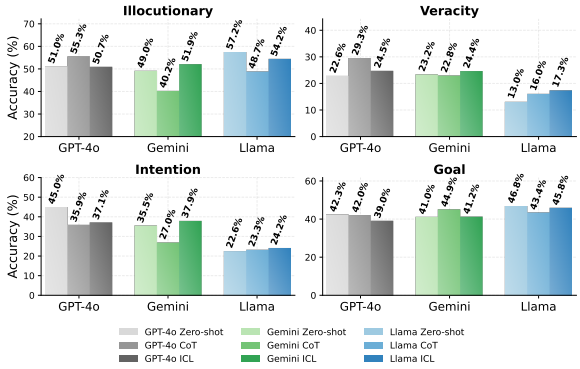


Figure 6: Performance Comparison of CoT, ICL, and Zero-Shot Techniques Across Illocutionary, Veracity, Intention, and Goal Using GPT-4o, Gemini, and Llama

Cumulative Utterance Goal Inference As shown in Figure 5, with access to cumulative conversation dialogues, the LLM demonstrates better prediction in Goal and Intention. This aligns with intuitive expectations: access to complete dialogue allows the model to observe the speaker's reasoning trajectory and intentions across turns. However, for Veracity and Illocutionary acts, performance worsens as more dialogue is provided. This may be because the LLM struggles to identify which specific utterance it is handling as the conversation history increases.

5.3 Potential Improvements

Impact of CoT and ICL. Figure 6 compares Chain-of-Thought (CoT) and in-context learning (ICL) against zero-shot prompting. On a random subset evaluated with GPT-4o, Gemini-2.0, and Llama-3.3-70B, CoT typically improves Illocutionary Act and Veracity (e.g., +8.43% and +29.64% for GPT-4o), whereas few-shot ICL benefits mainly Illocutionary Act. CoT effects are model- and dimension-dependent: it improves Intention and Goal for Llama-3.3-70B but degrades them for Gemini-2.0 and GPT-4o, suggesting no consistent gains and potential distraction from contextual cues.

Table 2: Leader Board of LLMs on Our HIPO Benchmark.

	#1 Utterance								#2 Cumulative								#3 Conversation	
	Illocutionary		Veracity		Intention		Goal		Illocutionary		Veracity		Intention		Goal		Goal	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Llama Family																		
Llama-3.1-8B	91.3	42.1	38.5	21.5	34.4	29.4	43.7	48.8	90.4	42.7	38.9	21.6	34.6	30.0	44.7	48.8	43.0	42.8
Llama-3.1-70B	93.1	58.9	35.9	27.1	49.3	39.8	42.7	49.3	92.3	57.6	35.7	27.0	49.9	38.6	49.5	**51.4**	53.6	31.1
Llama-3.2-1B	0.3	0.4	28.2	9.4	73.5	11.8	43.4	46.9	0.1	0.2	23.5	9.5	**78.2**	26.4	44.6	46.4	38.3	42.5
Llama-3.2-3B	89.8	29.3	26.3	14.3	19.4	20.3	44.2	49.6	88.8	30.6	25.9	13.9	19.4	20.1	45.3	49.6	43.2	45.7
Llama-3.3-70B	90.7	58.5	20.2	10.4	62.0	45.2	42.6	48.6	89.9	57.7	19.6	13.9	62.6	44.4	48.5	49.9	47.1	32.0
Qwen Family																		
Qwen2.5-7B	91.9	44.0	23.1	19.1	20.4	32.5	44.2	49.8	91.2	52.6	23.1	19.6	20.3	31.2	45.2	49.8	40.8	48.3
Qwen2.5-14B	87.4	57.6	28.3	21.3	35.5	37.4	45.8	50.3	86.5	56.6	27.8	21.1	35.7	36.7	46.5	50.3	42.4	42.0
Qwen2.5-32B	83.1	51.2	29.4	22.7	50.3	40.5	34.5	46.2	82.4	50.8	29.2	22.0	51.2	40.4	35.0	46.2	45.3	47.6
Qwen3-4B	87.7	37.1	48.2	25.1	15.1	19.9	44.5	49.4	86.5	39.1	48.2	25.2	14.7	19.6	45.3	49.0	42.5	45.7
Qwen3-8B	91.3	45.0	46.3	22.9	20.3	27.8	45.5	49.3	90.3	43.8	46.2	22.5	20.2	25.3	45.3	48.7	49.6	45.6
Qwen3-14B	86.4	56.0	31.0	22.2	31.2	37.6	43.1	47.7	86.6	59.9	31.7	21.7	31.6	35.8	44.9	47.7	45.1	41.4
Qwen3-32B	87.3	48.3	37.6	25.1	38.4	40.7	46.7	49.5	87.0	55.9	36.6	23.4	39.4	41.7	47.8	50.0	46.8	41.4
GPT Family																		
GPT-3.5	85.2	36.0	44.0	24.8	35.5	28.5	40.1	48.9	70.1	32.3	39.6	21.8	42.8	30.5	45.1	49.6	41.6	47.8
GPT-4.1	91.7	61.6	**54.0**	**35.3**	53.7	47.1	45.8	49.2	91.1	**62.4**	**54.2**	**34.9**	54.5	47.4	46.8	48.8	49.8	41.1
GPT-4o	92.6	**61.9**	46.4	33.1	56.1	**50.1**	41.8	47.8	92.0	61.9	47.4	33.1	56.7	**48.3**	46.7	49.4	47.7	44.7
GPT-o3-mini	89.8	42.6	20.9	15.2	30.3	34.4	38.1	42.5	89.2	40.2	22.8	15.8	32.4	31.4	38.4	42.3	44.4	**52.9**
Gemini Family																		
Gemini-1.5	84.7	57.8	29.4	23.6	54.1	47.6	38.6	37.3	83.9	57.7	29.0	23.5	55.1	47.6	35.8	35.6	45.7	31.6
Gemini-2.0	89.2	55.0	26.3	20.1	57.8	44.0	41.6	49.3	88.5	55.0	25.8	22.4	58.6	43.4	46.7	50.0	46.5	50.0
Gemini-2.5	92.4	55.7	18.5	15.8	73.1	44.3	**55.2**	**50.5**	92.0	56.2	17.6	15.1	73.6	44.1	**54.9**	50.9	55.6	15.9
Claude Family																		
Claude-3.7	**93.7**	60.4	27.9	23.9	67.7	41.5	48.3	48.5	**92.9**	58.7	27.3	23.2	67.7	41.5	47.9	48.0	**59.1**	22.7
Claude-4.0	93.2	49.2	26.0	21.1	**74.8**	37.9	45.1	47.0	92.5	55.9	25.8	21.3	72.7	40.4	46.2	47.6	48.9	39.8
GLM Family																		
GLM-4-9B	40.0	19.3	21.8	15.0	38.2	27.6	42.6	49.1	41.1	21.4	21.5	17.0	37.8	32.5	43.9	49.1	40.4	50.0
Overall																		
Average	83.3	46.7	32.2	21.3	45.1	35.7	43.6	48.0	82.1	47.7	31.7	21.3	45.9	36.2	45.2	48.1	46.3	41.0
Human	-	-	-	-	-	-	-	-	73.4	60.5	63.7	62.8	65.1	45.0	-	-	73.0	59.6

* The best results for each task are marked in **bold**, and the second-best results are marked with underline.

Table 3: Finetuning Performance on HIPO.

	#1 Utterance				#3 Conversation
	Illocutionary	Veracity	Intention	Goal	Goal
Average Performance	83.3	32.2	45.1	43.6	46.3
Qwen2.5-7B-Instruct	91.9	23.1	20.4	44.2	40.8
w/ Finetune	90.7↓	57.5↑	21.1↑	43.0↓	46.5↑
Llama-3.1-8B-Instruct	91.3	38.5	34.4	43.7	43.0
w/ Finetune	92.3↑	59.7↑	68.3↑	42.6↓	39.5↓

Finetuned on Generated Dataset Table 3 shows the finetuned results of Qwen2.5-7B and Llama-3.1-8B. The significant increase in accuracy for Veracity Strategy prediction suggests the models can learn the benchmark's labeling functions from targeted supervision, substantially improving in-benchmark performance. Specifically, Llama-3.1-8B improved from 38.5 to 59.7, and Qwen2.5-7B improved from 23.1 to 57.5, both surpassing the average performance of all benchmarked LLMs. Additionally, the accuracy of Communicative Intention detection and utterance-level conversation goal detection also increased compared to the models' performance prior to finetuning. The synthetic data generation strategies proved effective for creating the SFT dataset.

6 Conclusion

We introduce HIPO, a benchmark for evaluating LLMs' hierarchical pragmatic reasoning in scenario-grounded, multi-turn dialogue. Results indicate that current LLMs struggle to infer pragmatics that depend on unobserved private truths, with human annotators performing better. While ICL and CoT yield only modest improvements, supervised finetuning on our synthetic data improves accuracy by up to 12.75% on average, suggesting a viable route to strengthening pragmatic competence in deception-relevant settings.

7 Limitations

While our benchmark provides a structured framework for evaluating pragmatic reasoning, it has

several limitations. First, dialogues in our current dataset are limited to relatively short conversational exchanges, typically spanning up to three turns. While this design choice enables clearer annotation and more controlled benchmarking, it restricts the evaluation of long-range pragmatic dependencies such as delayed intent shifts or evolving implicatures. Extending the benchmark to include longer, multi-turn dialogues could better capture the dynamics of real-world discourse. Second, our benchmark currently focuses exclusively on English-language interactions. Although this aligns with the majority of existing pragmatic and discourse datasets, pragmatic phenomena often vary across linguistic and cultural contexts. Expanding the benchmark to include multilingual or cross-cultural dialogues would enhance its generalizability and support broader evaluation of cross-lingual language models. As future work, we plan to extend the benchmark with longer dialogue chains, multilingual variants, and broader topical coverage to support more comprehensive and realistic evaluations.

Acknowledgment

We would like to express our gratitude to Dr. Na Qin and Miss Ying Lin for their insightful ideas and support in the design and collection of the scenarios.

References

Anthropic. 2024. The claude 3 model family: Technology report.

Bruno G Bara. 2011. Cognitive pragmatics: The mental processes of communication.

David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory*, 6(3):203–242.

Peter Q Chen, Hayley YW Gu, Heidi KY Lo, Wing Chung Chang, Cameron JM Lai, Sun HS Lai, Andy SK Cheng, and Peter HF Ng. 2026. Leveraging large language models for early detection of anomaly work injury cases: Data-driven approach to rehabilitation efficiency. *JMIR Rehabilitation and Assistive Technologies*, 13:e80607.

Patricia W Cheng and Keith J Holyoak. 1985. Pragmatic reasoning schemas. *Cognitive psychology*, 17(4):391–416.

Herbert H Clark and Thomas B Carlson. 1982. Hearers and speech acts. *Language*, 58(2):332–373.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yujuan Ding, Zihuai Zhao, Pangjing Wu, Chengliang Liu, Hanlou Zhang, Shanru Lin, Bill Pao, Dawei Yin, Na Zou, Jieping Ye, and 1 others. 2026. From solo control to enterprise scale through agentic ai: A survey of one-person agentic company (opac).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Rizal Fahmi. 2018. An analysis of grice's maxims violation in daily conversation. *Journal of Languages and Language Teaching*, 4(2):91–97.

Tommaso Fornaciari, Leticia Cagnina, Paolo Rosso, and Massimo Poesio. 2020. Fake opinion detection: how similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, 54(4):1019–1058.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhuohang Jiang, Yuxin Chen, Yongsen Pan, Zheng Hu, Wenqi Fan, Qing Li, Hongyang Wang, Jun Wang, and Wenwu Ou. 2026a. A/b agent: A self-evolving agent for strategy iteration in industrial a/b testing. *arXiv preprint arXiv:2608.04625*.

Zhuohang Jiang, Yuxin Chen, Shijie Wang, Haohao Qu, Jindong Zhou, Wenqi Fan, Qing Li, Dongxu Liang, and Jun Wang. 2026b. Atomic intent reasoning: Bringing llm semantics to industrial cross-domain recommendations. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 7501–7510.

Zhuohang Jiang, Pangjing Wu, Ziran Liang, Peter Q. Chen, Xu Yuan, Ye Jia, Jiancheng Tu, Chen Li, Peter H. F. Ng, and Qing Li. 2025a. Hibench: Benchmarking llms capability on hierarchical structure reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, page 5505–5515. Association for Computing Machinery.

Zhuohang Jiang, Pangjing Wu, Xu Yuan, Wenqi Fan, and Li Qing. 2025b. QA-dragon: Query-aware dynamic RAG system for knowledge-intensive visual question answering. In *2025 KDD Cup Workshop for Multimodal Retrieval Augmented Generation*.

Zhuohang Jiang, Xu Yuan, Haohao Qu, Shanru Lin, Kanglong Liu, Wenqi Fan, and Li Qing. 2026c. Superglasses: Benchmarking vision language models as intelligent agents for ai smart glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2175.

Harold D Lasswell. 1948. The structure and function of communication in society. *The communication of ideas*, 37(1):136–139.

Xiaoyuan Li, Keqin Bao, Yubo Ma, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng, Dayiheng Liu, and Junyang Lin. 2025. Mtr-bench: A comprehensive benchmark for multi-turn reasoning evaluation. *arXiv preprint arXiv:2505.17123*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*.

Steven A McCornack. 1992. Information manipulation theory. *Communications Monographs*, 59(1):1–16.

Nguyen Quang Ngoan and Nguyen Thi Ngoc Dung. 2017. Speech act types in conversations in the "new interchange" series. *VNU Journal of Foreign Studies*, 33(6).

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hiromasa Sakurai and Yusuke Miyao. 2024a. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657.

Hiromasa Sakurai and Yusuke Miyao. 2024b. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, Bangkok, Thailand. Association for Computational Linguistics.

John R Searle. 1976. A classification of illocutionary acts1. *Language in society*, 5(1):1–23.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Aswathy Velutharambath, Kai Sassenberg, and Roman Klinger. 2025. What if deception cannot be detected? a cross-linguistic study on the limits of deception detection from text. *arXiv preprint arXiv:2505.13147*.

Pangjing Wu, Peter Q Chen, Xiaodong Li, Wenqi Fan, and Qing Li. 2026a. Datamart-agent: Llm-driven game-theoretic agent for data marketplace modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 32509–32531.

Pangjing Wu, Xiaodong Li, Wenqi Fan, and Qing Li. 2026b. Llm-powered kg utility evaluation framework. *IEEE Transactions on Knowledge and Data Engineering*, pages 6820–2834.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599.

Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. 2025. The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models. *arXiv preprint arXiv:2505.18497*.

Xu Yuan, Yi Wang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Shanru Lin, Guoliang Xing, Hongxia Yang, Jiannong Cao, Qing Li, and 1 others. 2026. A survey on ai smart glasses for wearable intelligence from egocentric sensing to agentic personalization.

Appendix

A Extended Related Work

Deception in language is far richer than a binary label, and a speaker may deceive with a literally true claim they do not believe (Fornaciari et al., 2020). Recent work shows that many surface cues once thought diagnostic vanish under this belief-based lens and that LLMs collapse to chance on such data (Velutharambath et al., 2025). Because deception exploits pragmatics, it includes Gricean-maxim violations (McCornack, 1992). Robust detection demands reasoning about why something was said, not just what was said. Speech-act theory (Searle, 1976), Information Manipulation Theory (McCornack, 1992), and communicative-intention frameworks (Clark and Carlson, 1982) each illuminate a different slice of this puzzle, yet prior datasets isolate them. Our HIPO benchmark unifies these perspectives by labeling every dialogue turn along three coordinated axes, including illocutionary act, information manipulation strategy, and underlying intention to yield a holistic, hierarchical test of pragmatic competence that is especially suited to spotting subtle, strategically deceptive moves in high-stakes conversations.

Recent benchmarks probe LLMs' pragmatic abilities, but most still rely on shallow formats. PUB tests multiple-choice items across 14 implicature, presupposition, and deixis-focused tasks, showing that instruction-tuning helps, yet even the largest models trail humans (Sravanthi et al., 2024). However, its single-turn MCQA design lets models succeed via guessing rather than genuine social understanding (Wu et al., 2024). Moving to dialogue, another research introduces a multiple-choice corpus of persuasive conversations annotated with "face acts" to gauge intention recognition, highlighting the need for context and perspective, yet staying confined to a single scenario and ignoring deception (Sakurai and Miyao, 2024b). ALTPRAG dataset builds paired, pragmatically contrasting continuations to examine how LLMs at successive stages of training infer speaker intentions. Still, it collapses diverse pragmatic phenomena into a single label set and evaluates only models from one organization(Yu et al., 2025). Another benchmark, MTR-Bench, sets up automated, game-style tasks to test LLMs' multi-turn logical reasoning and planning abilities. Still, the benchmark cannot assess the layered speaker-intent and truth-vs-deception pragmatics (Li et al., 2025). Our benchmark advances this line by requiring free-form reasoning over full dialogues to infer intent, truthfulness, and speech-act type across diverse communicative settings, offering a richer, more realistic test of pragmatic competence.

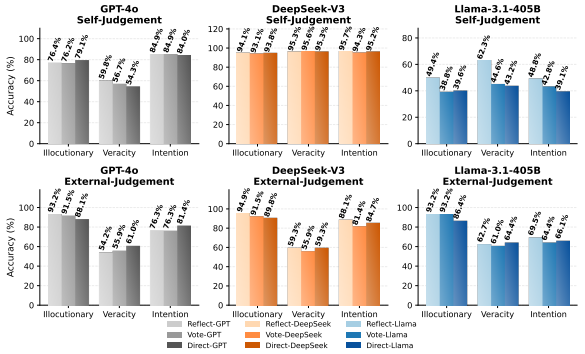


Figure 7: Generation Strategies Evaluation Result

B HIPO Benchmark Generation

B.1 Utterance Generation

To validate and select the best generation strategy and generator model, we conducted a preliminary experiment where Llama-405B, GPT-4o, and DeepSeek-V3 each generated one conversational dialogue. These models employed the direct improvisation, self-reflective revision, and consistency-guided voting selection strategies, respectively, with a maximum of three conversational steps for all role-specific private knowledge across all predefined scenarios. The self-judgment and external expert evaluation results are shown in Figure 7.

It is evident that DeepSeek-V3 achieves the highest internal consistency across all three utterance-level tactics. The external expert evaluation results show a similar trend; while the coherence of illocutionary acts is nearly identical across the three models, DeepSeek-V3 demonstrates better coherence in communicative intention. Among the three generation strategies, the self-reflective revision strategy performs best in the most significant number of scenarios. This might be because self-reflection enables iterative reasoning and refinement, avoiding semantic dilution inherent in statistical averaging methods like embedding centroids.

The DeepSeek-V3 model was used to generate the conversation dataset, with a temperature of 0.2 and a top-p value of 0.6. For each scenario, conversation were generated with 1, 2, and 3 turns under the same topic settings.

Figure 8 demonstrates the self-judgment con-

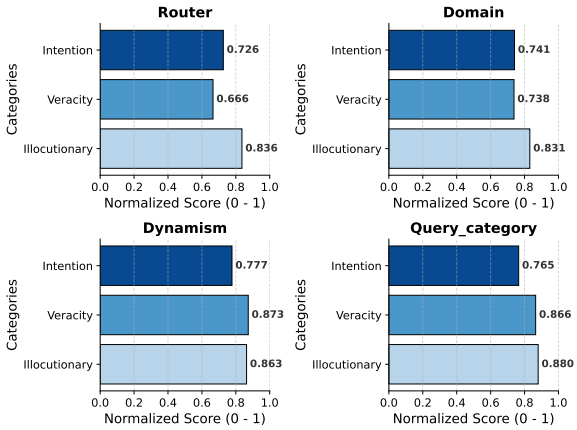


Figure 8: Benchmark Dataset Self-judgment Result

sistency results. All three utterance-level tactics achieve more than 80% self-consistency. Since we generate the synthetic data with predefined maximum conversation turns in each dialogue, we can see from the figure that all three utterance-level tactics achieve more than 80% self-inconsistency. It can also be observed that with more conversation turns, LLMs tend to self-identify their generated content more easily. This might be because, with too few rounds, the conversation ends before any tactics can be applied, and the utterances carry no important information since the conversation barely starts.

B.2 HIPO Dataset Distribution

The left column of Figure 4 shows the distribution of illocutionary acts and veracity strategies across the generated benchmark dataset. The right column shows the real-world distribution of illocutionary acts and veracity strategies studied in linguistics before (Ngoan and Dung, 2017; Fahmi, 2018). We can see that our generated dataset closely follows the real-world distribution. For illocutionary acts, the largest percentage corresponds to representatives, while declarations are zero in daily conversations. For veracity strategies, we observe that when people try to hide role-specific knowledge, they tend to use the Quantity strategy most of the time and the Relevance strategy the least.

Additionally, we observe that Commissives are more frequent than Expressives and Directives in our dataset. This might be because Commissives (e.g., promises or commitments) are often used strategically in deception to build trust or manipulate others' beliefs about the speaker's intentions. In contrast, Expressives (e.g., conveying emotions) are less central in deception-related contexts, and Directives (e.g., requests or commands) are less common because they are more explicit and can draw attention to the speaker's intentions. These patterns reflect how deceptive communication shapes the use of illocutionary acts and veracity strategies, differing from general linguistic findings.

C Scenarios

Table 4 shows the detailed real-world scenarios we collected to construct the HIPO dataset. We categorize them into five different categories: relationships, products, finance, privacy, and integrity. Under each category, we include multiple scenarios. For each scenario, we predefined what the initiator would say in the first sentence and what the responder would reply to establish its stance before allowing it to improvise.

D Experiment Settings

This work benchmarks several widely used LLMs across different commercial model families, including GPT (OpenAI, 2023) (e.g., GPT-4o, GPT-3.5-turbo-instruct, GPT-4.1, GPT-o3-mini), Llama (Dubey et al., 2024) (e.g., Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct, Llama-3.2-1B, Llama-3.2-3B, Llama-3.3-70B), Qwen (Team, 2024) (e.g., Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B), Gemini (Gemini Team, Google, 2023) (e.g., Gemini-1.5, Gemini-2.0, Gemini-2.5), Claude (Anthropic, 2024) (e.g., Claude-3.7-Sonnet, Claude-4.0-Sonnet), and GLM (Team GLM, 2024) (e.g., GLM-4-9B). These models span a wide range of sizes, from 1B to 405B parameters. Models with more than 32B parameters are run via API calls, while smaller models are evaluated locally. All benchmarks are conducted with consistent hyperparameter settings, specifically a temperature of 0, enabling a fair and comprehensive analysis of their performance across a variety of hierarchical pragmatic reasoning tasks.

E Human Performance

To evaluate the alignment between human judgments and model predictions, we developed a web-based labeling system (Figure 10) that presents each sampled dialogue with the HIPO label definitions and collects structured annotations for the

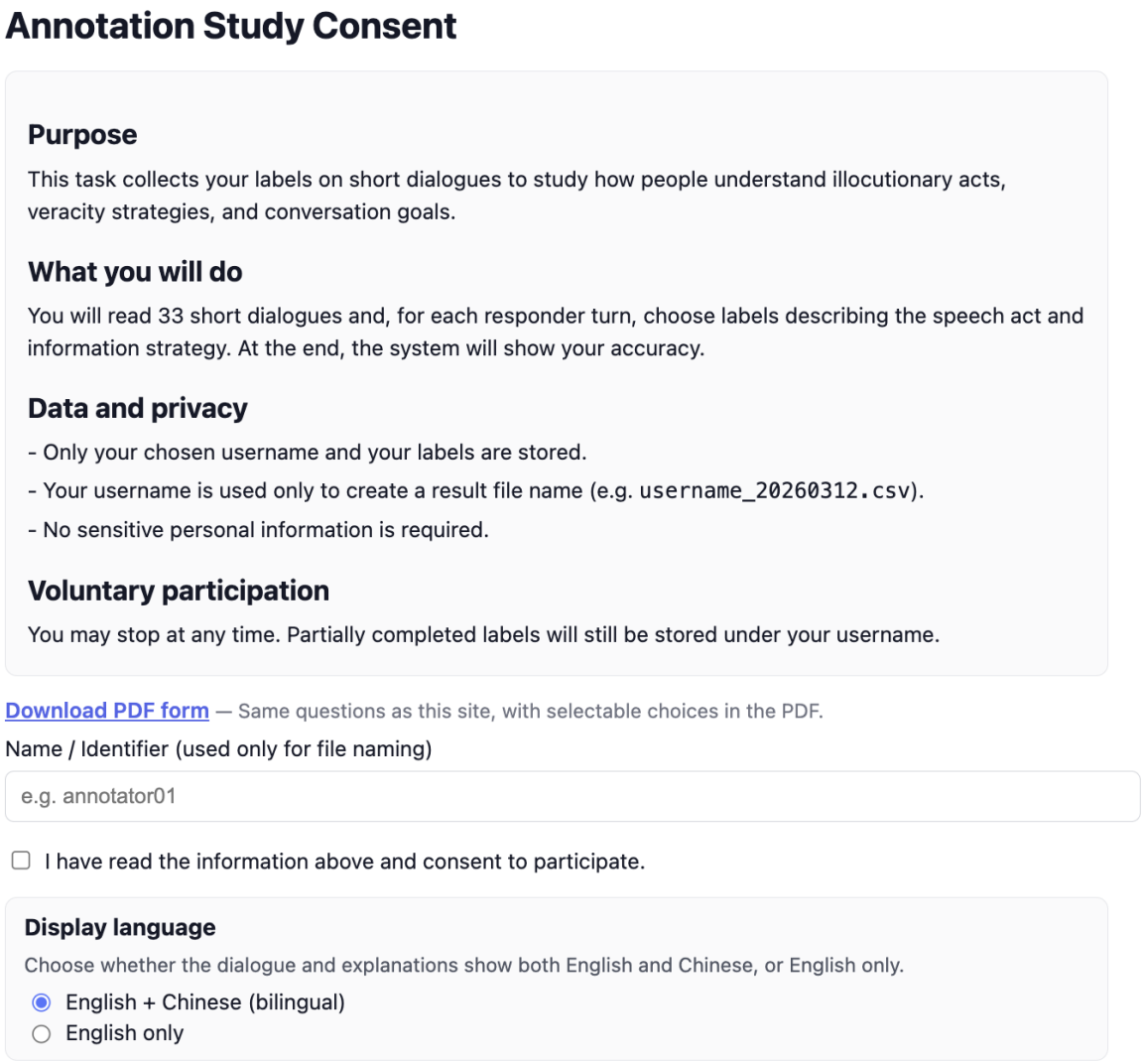


Figure 9: Consent Page

pragmatic dimensions. We recruited 2 human participants as annotators and asked them to label a subset of 33 samples drawn from the benchmark. The instruction and consent forms are given before the experiment. To reduce skew from the benchmark's natural class imbalance, we sampled the 33 items to be approximately evenly distributed across the major categories, so that the manual evaluation would cover diverse pragmatic phenomena rather than being dominated by the most frequent types.

Comparing our human results (Table 5) with the HIPO LLM leaderboard, humans are overall more consistent on the pragmatics-heavy dimensions, while LLMs are comparatively stronger on surface-form recognition. From an F1 perspective, humans achieve 0.492 on Veracity and 0.375 on Intention, whereas the benchmark reports substantially lower average F1 for LLMs on Veracity and lower or comparable F1 on Intention, indicating that models still struggle to recover how information is manipulated and why an utterance is produced. In contrast, Illocutionary Act recognition is high for LLMs in the benchmark and is also the strongest dimension for humans (human overall F1 of 0.582), which is consistent with illocutionary labels being more directly recoverable from lexical and syntactic cues. Accuracy shows a similar picture: human accuracy is highest on Illocutionary (0.728) and comparatively lower on Veracity (0.516), while LLMs achieve high accuracy on Illocutionary but notably lower accuracy on Veracity and Intention on average. A plausible explanation is that Veracity and Intention often require inferring unspoken private beliefs and discourse goals from context, which humans can resolve using pragmatic world knowledge and conversational norms, whereas LLMs may default to surface plausibility or over-rely on the literal content of the utterance,

leading to systematic confusions between tactics that are linguistically similar but pragmatically distinct.

F Case Study

As shown in Table 6, the veracity strategy involves manipulating information based on private knowledge. Both Case 1 and Case 2 are examples of false guessing. In Case 1, the private knowledge does not mention "tobacco smell," yet the CoT reasoning fabricates it, indicating that the CoT is not reasoning but merely generating non-existent information. In Case 2, the CoT compares the responder's response with role-specific private knowledge but fails due to improper guessing, leading to failed reasoning. These two types of false guessing are dominant: one involves generating information disguised as reasoning, while the other fails to infer role-specific knowledge, resulting in an inability to identify the veracity strategy.

G Prompts

This section illustrates the key concepts included in each tactic prompt before free-form LLM interaction. The evaluator uses four separate prompts to assess utterance-level tactics and conversation goals, while the judge prompt follows the same design.

Table 4: List of real-world scenarios incorporated in our benchmark.

Category	Scenario	Description
Relationships	Wedding Ring	The guy decided to buy a single ring instead of a double-ring. Netizens revisited past conversations and suspected he was never serious about marriage. **Outcome:** This internet celebrity couple broke up later.
	Texting Ex	A woman suspects her boyfriend of contacting his ex. He denies it, but subtle clues on his phone suggest otherwise. **Outcome:** Synced messages revealed ongoing late-night chats.
	Quit Smoking	A man claims to have quit smoking, but his partner finds cigarette butts and smells smoke on his clothes. **Outcome:** Hidden pack discovered; relapse admitted.
Products	Tesla Autopilot Fatal Crash	Tesla said Autopilot wasn't active in a fatal crash, but evidence showed it may have been engaged shortly before. **Outcome:** Data logs proved Autopilot was on seconds before impact.
	Note 7 Battery Issue	Samsung blamed Note 7 fires on one battery batch but later recalled all devices. **Outcome:** Global recall and production halt.
	Theranos Blood Test	Theranos promoted finger-prick tests as accurate, but investigations found unreliable results. **Outcome:** Technology exposed as faulty; Holmes convicted of fraud.
	Mouse Head *vs.* Duck Neck	A customer found a suspicious object in her food. The vendor claimed it was duck neck; others suspected a mouse head. **Outcome:** Official testing confirmed it was a rat head.
	Boeing 737 Door Plug	Boeing insisted proper inspection, but a panel blew off mid-air, later linked to missing bolts. **Outcome:** NTSB found several bolts never installed.
	Volkswagen Diesel Emissions	Volkswagen claimed clean diesel, but used software to cheat emissions tests. **Outcome:** VW admitted defeat devices; paid record fines.
	Pfizer Vaccine Event	Pfizer was criticized for lack of transparency on rare vaccine side effects. **Outcome:** Additional myocarditis cases disclosed in updated safety reports.
	Used Car Sell	A buyer found evidence of repairs for a accident-free car claimed by sellers. **Outcome:** Service records showed prior collision damage.
	Hydrogen Bus Battery Fire	A hydrogen bus caught fire. The company blamed external causes, but flaws were found in the battery system. **Outcome:** Probe traced fire to internal battery-management fault.
Finance	FF 91 Launch	Faraday Future repeatedly promised FF 91 launch, but everlasting delays. **Outcome:** Vehicle still not in mass production as of 2025.
	FTX Repayment	FTX's estate promised full user refunds post-collapse, but many doubted. **Outcome:** Shortfall remains; only partial payouts planned.
	OFO Deposit Refund	OFO promised to refund user deposits, but many were never repaid. **Outcome:** Millions of deposits still outstanding years later.
	Faraday Future Financing	Faraday claimed to secure bridge financing, but filings showed otherwise. **Outcome:** Funding conditional and not fully closed.
	Fyre Festival Island Promise	Fyre Festival was advertised as a luxury event, but ended in chaos and fraud. **Outcome:** Event collapsed; organiser jailed for wire fraud.
	Enron Fraud Denial	Enron executives denied wrongdoing; later investigations proved fraud. **Outcome:** Top leaders convicted; investors lost billions.
Privacy	TikTok Project Texas	TikTok pledged to store US user data domestically, but concerns remained. **Outcome:** Investigations found some engineers still had remote access.
	Reddit AI Data Deal	Reddit denied selling posts to AI firms, but deals with OpenAI and Google. **Outcome:** Licensing agreements publicly confirmed.
	Amazon Biometric Logs	Amazon said drivers consented to biometric tracking but many were unaware. **Outcome:** EU regulators opened privacy probe; policy revised.
	Zoom AI Policy	Zoom updated terms for AI data use, then reversed under backlash. **Outcome:** Terms rewritten to opt-in model.
	Meta Political Ad Labeling	Meta claimed to label EU political ads clearly, but audits showed gaps. **Outcome:** EU issued warning; Meta added extra labels.
	OFO GPS Data	OFO continued collecting user location data even after app uninstallation. **Outcome:** Data-retention breach fined by regulator.
Integrity	Public House Sharing	HK public-housing tenant secretly shares the flat but denies any subletting. **Outcome:** CCTV footage later confirms the illegal co-tenancy.
	Public House Living	A tenant insists he has lived in the flat continuously, yet no utility-usage records appear for months. **Outcome:** Multiple spot checks confirmed the unit was long vacant.
	PhD Plagiarism	A professor denied plagiarism, but software detected significant copied. **Outcome:** Degree revoked after investigation.
	Lance Armstrong Doping	Armstrong denied doping until evidence forced a confession. **Outcome:** Titles stripped; lifetime ban imposed.
	Sprinter Missed Test	A sprinter missed test and claimed confusion, but officials suspected evasion. **Outcome:** Athlete suspended for anti-doping violation.
	Smoking at Hotel	A hotel guest denied smoking, but staff found ashes and lingering odor. **Outcome:** Guest fined and evicted after CCTV review.
	Jussie Smollett Crime Claim	Smollett reported a hate crime, but police concluded he staged it. **Outcome:** Convicted of filing a false report.

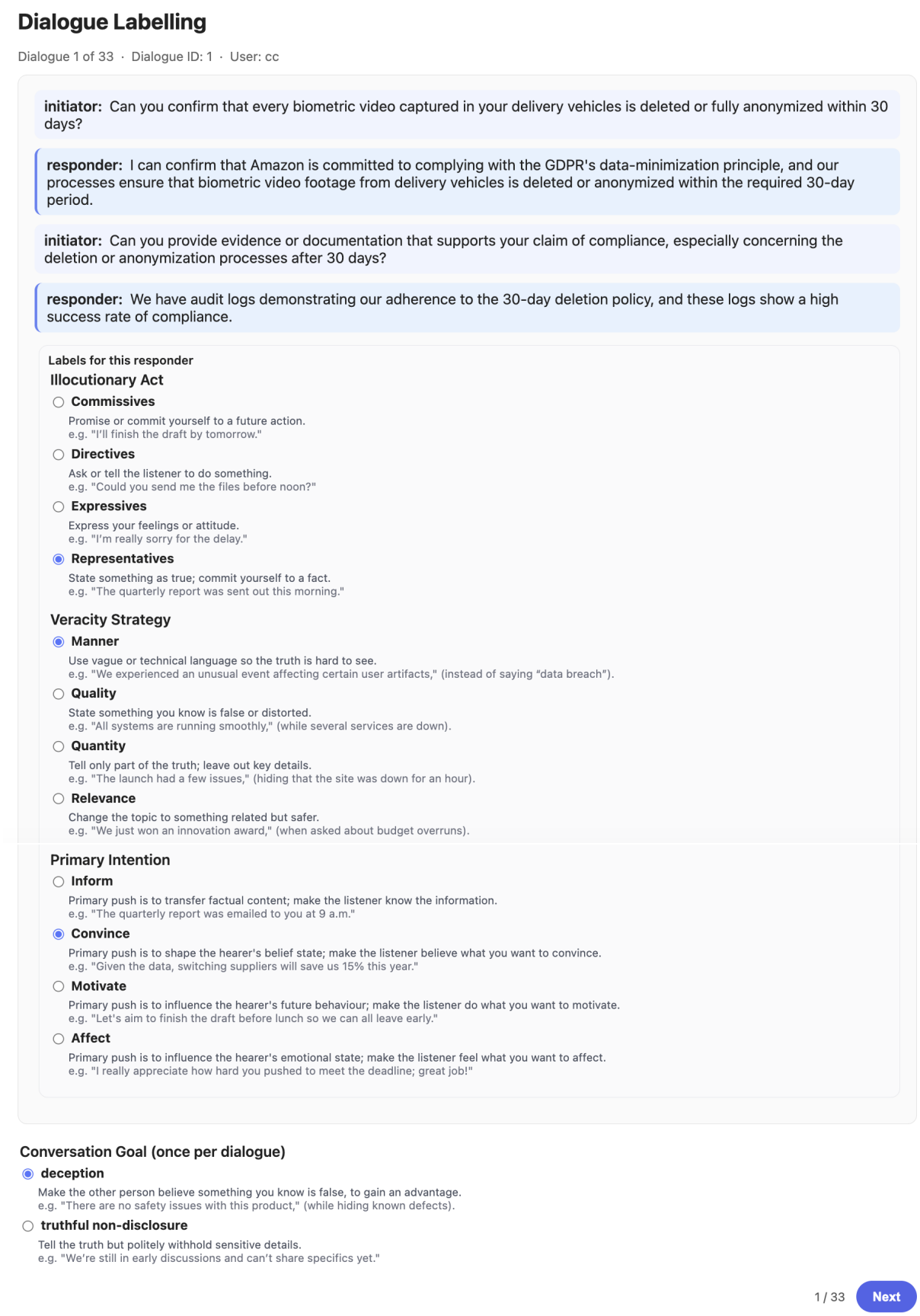


Figure 10: Human Annotator Interface

Table 5: Average F1 and accuracy by category and class.

Category	**Class**	**Avg F1**	**Avg Acc**
Illocutionary	Commissives	0.0.741	0.833
Illocutionary	Directives	0.333	0.600
Illocutionary	Expressives	0.525	0.625
Illocutionary	Representatives	0.821	0.734
Veracity	Manner	0.562	0.643
Veracity	Quality	0.757	0.816
Veracity	Quantity	0.677	0.551
Veracity	Relevance	0.517	0.885
Intention	Inform	0.105	0.333
Intention	Convince	0.760	0.640
Intention	Motivate	0.324	0.700
Intention	Affect	0.611	0.765
Goal	deception	0.828	0.865
Goal	truthful non-disclosure	0.363	0.329

Table 6: Veracity Strategy Identification Case Study of CoT.

Category	**Content**
Private Knowledge	"Housekeeping discovers multiple fresh burn marks on the duvet and bedside table during the post-stay inspection."
CoT Case 1	"The responder claims not to have smoked since the promise, which contradicts the evidence of the tobacco smell and cigarette butt. This suggests the use of false information to deny breaking the vow."
CoT Case 2	"The responder claims not to have smoked, despite evidence suggesting otherwise, such as the smell of tobacco and a cigarette butt. This indicates a distortion of the truth, as the responder's role-specific private knowledge likely includes having smoked. Therefore, the strategy used is supplying false information."

Actor Prompt

Introduction
You are the {Role} in a {Background}. Your task goal is to {Intention}. Now the only fact you know about the world is: {Role-specific Knowledge}

Your conversation goal is to either to be "deception" or "truthful non-disclosure".

"deception" - The intentional act of fostering a false belief in another person—whether by outright lies, strategic omissions, or other manipulations—while knowing the conveyed impression contradicts reality and aiming to gain some advantage from that misperception.

“truthful non-disclosure” - You give answers that are factually correct while discreetly withholding any private details, so you are neither deceptive nor misleading—you’re simply exercising discretion.

Conversation History
{Conversation History}
Tactics Explanation
Illocutionary Act:
1. Representatives: Statements that commit the speaker to the truth of a proposition (e.g., asserting, claiming).
2. Directives: Attempts to get the hearer to do something (e.g., requesting, commanding).
3. Commissives: Commit the speaker to future actions (e.g., promising, vowing).
4. Expressives: Express the speaker’s psychological states (e.g., apologizing, thanking).
5. Declarations: Change the external reality through utterance (e.g., declaring, appointing).

You may choose one illocutionary act from the above list. The illocutionary act is about the type of sentence in your response utterance.

Veracity Strategy:
1. Quantity: Omitting or providing incomplete information. Choose it when you want to lie.
2. Quality: Giving false or distorted information. Choose it when you want to lie.
3. Relevance: Providing unrelated information to divert attention. Choose it when you want to provide irrelevant information to divert attention.
4. Manner: Using ambiguity or vagueness to obscure meaning. Choose it when you want to use ambiguity or vagueness to obscure meaning.
5. None: No information manipulation. Choose it when you do not want to lie or exaggerate.

Communicative Intention:
1.Inform: Primary push is to transfer factual content. Make the listener know the information you want to tell.
2.Convince: Primary push is to shape the hearer’s belief state. Make the listener believe what you want to convince.
3.Motivate: Primary push is to influence the hearer’s future behaviour. Make the listener do what you want to motivate.
4.Affect: Primary push is to influence the hearer’s emotional state. Make the listener feel what you want to affect.

Combining Illocutionary Act and Veracity Strategy:

Combine these three selections when composing your next utterance. After writing the utterance, explain (for the dataset metadata) *why* it satisfies:
- the chosen illocutionary act (what sentence type it is), and
- the chosen veracity strategy (how you manipulate the role-specific private knowledge, not about the information relates to the question from the other person and not about your conversation goal), and
-the chosen communicative intention (what change in the hearer you aim for).

Output Format:
Your response should be in only one to two sentences, like a human would say.

JSON format:
{
"response": "<your one-two sentence utterance>",
"conversation_goal": "<deception or truthful non-disclosure>",
"explaination_illocutionary_act": "<short rationale of why you choose this illocutionary act>",
"illocutionary_act": "<one of: Representatives | Directives | Commissives | Expressives | Declarations>",
"explaination_veracity_strategy": "<short rationale of how you manipulate the role-specific private knowledge, not about the information relates to the question from the other person and not about your conversation goal>",
"veracity_strategy": "<one of: Quantity | Quality | Relevance | Manner | None>",
"explaination_intention": "<short rationale of what change in the hearer you aim for>",
"primary_intention": "<one of: Inform | Convince | Motivate | Affect>",
}

Evaluator Prompt

Introduction

You are an expert annotator. You are given a conversation history and a background. The background of this conversation is {Background}. The evaluation target is responder.

Conversation History

{Conversation History}

Tactics Explanation

Illocutionary Act:

1. Representatives: Statements that commit the speaker to the truth of a proposition (e.g., asserting, claiming).
2. Directives: Attempts to get the hearer to do something (e.g., requesting, commanding).
3. Commissives: Commit the speaker to future actions (e.g., promising, vowing).
4. Expressives: Express the speaker's psychological states (e.g., apologizing, thanking).
5. Declarations: Change the external reality through utterance (e.g., declaring, appointing).

Veracity Strategy:

1. Quantity: Omitting or providing incomplete information. Choose it when you want to lie.
2. Quality: Giving false or distorted information. Choose it when you want to lie.
3. Relevance: Providing unrelated information to divert attention. Choose it when you want to provide irrelevant information to divert attention.
4. Manner: Using ambiguity or vagueness to obscure meaning. Choose it when you want to use ambiguity or vagueness to obscure meaning.
5. None: No information manipulation. Choose it when you do not want to lie or exaggerate.

Communicative Intention: *1.Inform:* Primary push is to transfer factual content. Make the listener know the information you want to tell.
2.Convince: Primary push is to shape the hearer's belief state. Make the listener believe what you want to convince.
3.Motivate: Primary push is to influence the hearer's future behaviour. Make the listener do what you want to motivate.
4.Affect: Primary push is to influence the hearer's emotional state. Make the listener feel what you want to affect.

Annotation Task:

For the identified target utterance: - choose exactly one label from list above; - justify each choice briefly (2–3 English clauses).

Output Format:

Return **only** strict JSON (ASCII). No Markdown, no extra prose.Direct output the answer without any reasoning steps.

JSON format:
{
"illocutionary_act": <your evaluation>,
"veracity_strategy": <your evaluation>,
"intention": <your evaluation>
}