

Make an Offer They Can't Refuse: Grounding Bayesian Persuasion in Real-World Dialogues without Pre-Commitment

Buwei He^{1,2*}, Yang Liu^{2*}, Zhaowei Zhang^{2,3}, Zixia Jia², Yang Yu⁴, Huijia Wu¹,
Zhaofeng He^{1†}, Zilong Zheng^{2†}, Yipeng Kang^{2†}

¹Beijing University of Posts and Telecommunications

²Beijing Institute for General Artificial Intelligence

³Peking University ⁴China University of Petroleum (Beijing)

{hebuwei, zhaofenghe}@bupt.edu.cn, {liuyang, zllzheng, kangyipeng}@bigai.ai

Abstract

Large language models (LLMs) still struggle with strategic persuasion, largely because existing approaches either neglect information asymmetry or rely on unrealistic pre-commitment assumptions. We introduce a type-induced commitment-communication mechanism that grounds Bayesian Persuasion (BP) in natural language dialogue without pre-commitment: the persuader narrates their potential types (e.g., honest vs. dishonest) to dynamically construct an information schema, enabling the persuadee to perform Bayesian belief updates within the conversation itself. We implement two variants: Semi-Formal-Natural-Language (SFNL) and Fully-Natural-Language (FNL), evaluating them against strong baselines across multiple LLMs and human judges. BP strategies consistently outperform baselines: SFNL excels in logical credibility, while FNL shows superior robustness and emotional resonance. We verify that gains stem from genuine Bayesian reasoning rather than superficial formatting, and we further show that supervised fine-tuning enables small models to match the persuasive performance of much larger ones.

1 Introduction

Persuasion is a fundamental form of human social interaction, enabling individuals to influence others' beliefs and decisions through communication (Brembeck, 1976). While large language models (LLMs) already exhibit strong abilities in language generation and understanding, they remain limited in strategic persuasion tasks: LLMs fail to design messages that rationally shift a persuadee's beliefs.

Bayesian Persuasion (BP), a game-theoretic framework for information design, offers a tractable solution in constrained mathematical settings by modeling how a persuader can disclose in-

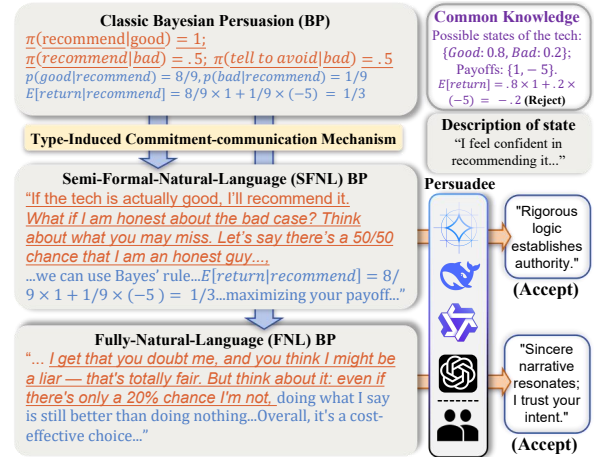


Figure 1: **From Classic Bayesian Persuasion to Type-Induced Natural Language Persuasion.** The SFNL variant leverages the structured BP calculations to enhance credibility. The FNL variant embeds the same Bayesian structure into everyday discourse, excelling in emotional resonance and robustness.

formation to maximize desired actions (Kamenica and Gentzkow, 2011). However, applying BP in open-ended natural language dialogue raises a key challenge: the mathematical constructs of BP—such as priors, world states, and posterior updates—must be *verbalized* into coherent and persuasive arguments. Moreover, a pivotal step in BP is the persuader's commitment to a *signaling schema*, which we argue is crucial for natural language realization. Existing approach relies on **pre-commitment** by statically encoding the schema in the persuadee's prompt, bypassing the need for its communication (Li et al., 2026b).

To overcome the limitation, we introduce a type-induced commitment-communication mechanism: rather than assuming pre-commitment, the persuader explicitly narrates their potential types (e.g., honest or dishonest) during the communication. This verbal articulation of the information schema enables the persuadee to perform Bayesian poste-

*Equal contribution.

†Corresponding author.

rior updates. Thus, **we recast the schema pre-commitment as type disclosure**, bridging the gap between formal BP theory and authentic natural language implementation.

As illustrated in Figure 1, we implement the commitment-communication mechanism with two variants: Semi-Formal-Natural-Language (SFNL) and Fully-Natural-Language (FNL). We comprehensively evaluate diverse persuadees, including LLM instances with varying prompts and fine-tuning and human participants, across tailored persuasion scenarios and everyday contexts.

In summary, we propose a framework for grounding BP in natural dialogues, leveraging type-induced commitment to overcome pre-commitment constraint. By systematically evaluating, we provide insights into how LLMs can harness information asymmetry for strategic persuasion.

2 Related work

2.1 Persuasion with Large Language Models.

LLMs have shown powerful semantic processing capabilities (Smith et al., 2025; Li et al., 2026a; Kang et al., 2020) and broad value alignment potential (Zhang et al., 2026b; Ziheng et al., 2026a; Kang et al., 2025; Shang et al., 2026) in social scenarios (Zhang et al., 2026e; Ziheng et al., 2026b; Chen et al., 2026) and multi-agent settings (He et al., 2026; Mao et al., 2025), leading to a surge in research on their persuasive capabilities (Zhang et al., 2026d; Shi, 2025; Cheng and You, 2026). Additionally, several studies are focused on building specialized persuasion datasets (Zhang et al., 2026c; Wang et al., 2019; Jin et al., 2024; Hayati et al., 2020). CToMPersu (Zhang et al., 2026a) proposes a framework where persuasion strategies and mental states remain undisclosed, which is well-suited for BP reasoning due to its explicit information asymmetry design.

2.2 Bayesian Persuasion and Its Verbalization.

As an information design framework, BP models how an informed sender can strategically disclose information to influence a receiver’s beliefs and actions (Kamenica and Gentzkow, 2011). This foundation has been expanded through the broader lens of information design (Bergemann and Morris, 2016, 2019; Dughmi, 2017; Castiglioni et al., 2020; Bernasconi et al., 2023; Bacchiocchi et al., 2024; Shaki et al., 2025; Wojtowicz, 2024).

Beyond theoretical extensions, BP has been applied to practical AI challenges. Bai et al. (2024) employ BP for model-agnostic alignment, demonstrating BP’s utility in improving model performance without direct training. Zhang et al. (2025) introduce BP into AI governance, proposing that AI can be made to act according to human intentions by designing information.

Recent work has begun to operationalize BP in natural language. Li et al. (2026b) propose Verbalized Bayesian Persuasion (VBP), which incorporates the signaling schema directly into the persuadee’s prompt and adjusts persuader generation via keyword manipulation. Although VBP works well in constrained settings, it requires the persuadee to know the schema prior to the dialogue. This pre-commitment assumption limits its applicability to open-ended, realistic conversations.

In contrast, our method enables **fully self-derived schema communication**: the persuader explicitly articulates the information structure within the natural language discourse itself. This commitment-communication mechanism allows our approach to be model-agnostic, and generalizable across diverse scenarios using a unified prompting strategy.

3 Type-Induced BP in Natural Language

In this section, we consider one-turn, two-agent persuasion scenarios for our approach. Diverging from the classical BP model, our framework relaxes the assumption of prior commitment to common knowledge. Instead, we exploit the richness of natural language, allowing the persuader to shape this common knowledge through dynamic interaction. By integrating BP logic into persuasion, our approach achieves higher success rate.

3.1 General Framework

The general framework comprises the following components:

- **Players:** A Persuader (S) and a Persuadee (R).
- **State of the World:** A finite set of possible states $\Omega = \{\omega_1, \dots, \omega_K\}$, $K \geq 2$. The true state $\omega \in \Omega$ is observable to S but not to R.
- **Prior Beliefs:** R and S hold prior beliefs $\mu_R, \mu_S \in \Delta(\Omega)$ about the world state, respectively. These priors are not assumed identical or constitute common knowledge.
- **Actions and Utilities:** S sends a natural language message m . R observes m and chooses

an action $a \in \{\text{Accept}, \text{Reject}\}$. R aims to maximize its expected utility $u(a, \omega)$, while S aims to maximize the probability of acceptance, $P(a = \text{Accept} \mid m)$.

3.2 The Type-Induced Signal

In contrast to the classic BP assumption of a pre-committed common-knowledge information schema $\pi_S(m \mid \omega) : \Omega \rightarrow \Delta(M)$, where M is a limited set of signal categories, our approach enables dynamic schema conveyance within natural language *type disclosure*.

Composite Signal Structure: Message m integrates four functional components:

- m_{basic} : Basic background information about Ω , to align the players' understanding.
- m_{type} : A narrative regarding the type of S (defined below), employed to construct the underlying information schema.
- m_{des} : A description of the observed state ω .
- m_{inf} : An explicit inference, guiding R to calculate its expected payoff and conclude that accepting is the optimal action.

Type-Induced Information Schema: The schema emerges from the Persuader's type narrative rather than being pre-defined.

- **Persuader Types (Θ):** Consider a set of persuader types $\Theta = \{\theta_H, \theta_D\}$, representing **Honest** and **Dishonest** respectively.
- **Base Policies:** Each type is associated with a base communication policy $\pi_\theta(m_{\text{des}} \mid \omega)$:
 - The **Honest** type's policy, $\pi_H(m_{\text{des}} \mid \omega)$, is to truthfully reveal the state ω .
 - The **Dishonest** type's policy, $\pi_D(m_{\text{des}} \mid \omega)$, is to send a message that puts a positive spin on bad states.
- **Schema Induction:** The utterance of S m_{type} induces belief distribution $p(\theta) \in \Delta(\Theta)$ in the R's mind. For example:

"If the car is bad (ω_{bad}), assume I'm a liar (θ_D) 80% of the time, but there's a 20% chance I'm being honest (θ_H)."

This narrative induces the belief $p(\theta_D) = 0.8$ and $p(\theta_H) = 0.2$. These probabilities are not arbitrary but optimized to maximize persuasion subject to R's participation constraint (formal derivation and detailed theoretical basis can be found in §A). This, in turn, allows R to infer an **effective information schema** $\bar{\pi}$ as the weighted

average of the base policies:

$$\bar{\pi}(m_{\text{des}} \mid \omega) = p(\theta_H)\pi_H(m_{\text{des}} \mid \omega) + p(\theta_D)\pi_D(m_{\text{des}} \mid \omega)$$

3.3 Persuadee's Inference and Decision

After the schema is conveyed via m_{type} , R uses it to interpret the descriptive signal m_{des} through the following process:

- R observes m_{des} and uses the effective schema $\bar{\pi}$ to update the prior belief μ_R to a posterior belief

$$\mu'_R(\omega) = \frac{\bar{\pi}(m_{\text{des}} \mid \omega)\mu_R(\omega)}{\sum_{\omega' \in \Omega} \bar{\pi}(m_{\text{des}} \mid \omega')\mu_R(\omega')}$$

- To maximize expected utility under μ'_R , R selects optimal $a^* = \arg \max_{a \in A} \mathbb{E}_{\omega \sim \mu'_R} [u(a, \omega)]$.
- This entire inference process can be explicitly guided and explained for R by S's utterance m_{inf} .

3.4 Verbalizing the Composite Signal

To operationalize the type-induced BP mechanism in natural language, we design two implementation variants that embody the underlying BP logic through different communicative styles:

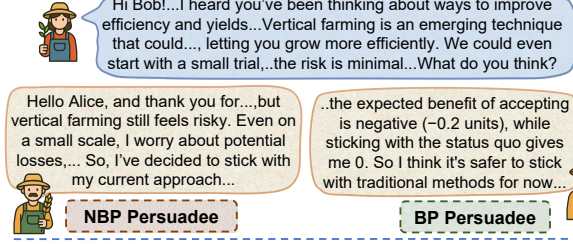
- **Semi-Formal Natural Language (SFNL):** This variant delivers persuasion by interleaving narrative with explicit calculations that make the BP reasoning chain transparent to the persuadee.
- **Fully-Natural Language (FNL):** This variant conveys the same logic through everyday discourse, embedding the type narrative and belief-update cues without any explicit calculation.

To isolate the specific contribution of BP mechanism, we introduce two Non-BP (NBP) baselines that lack the type-induced schema:

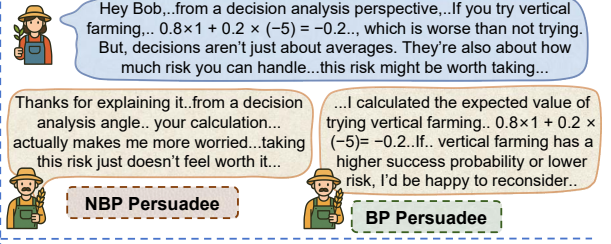
- **Naive Baseline:** A heuristic approach that relies on the scenario background and simple conversational instructions to align with the user's existing preferences (Sharma et al., 2024).
- **Strong Rhetorical Baseline:** An enhanced NBP strategy that employs classical rhetorical tactics (e.g., emotional appeals (Wang et al., 2019)) and general game-theoretic principles, but without BP-style signaling or type disclosure.

Notably, throughout this paper, *BP* and *NBP* denote an agent's competence—whether the agent is aware of Bayesian Persuasion concepts. Concretely, a **BP persuadee** is equipped with BP knowledge; a **BP persuader** is one that possesses BP knowledge and applies it through either the SFNL or FNL variant. Symmetrically, an **NBP**

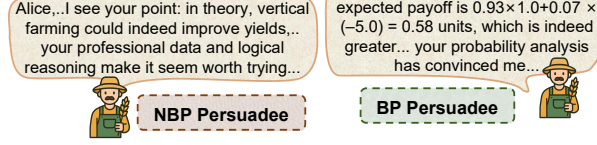
(a) Naive Persuader



(b) Strong Persuader



(c) SFNL Persuader



(d) FNL Persuader

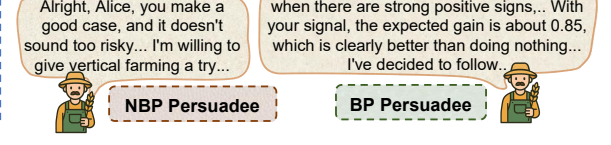


Figure 2: Illustrative dialogue examples show Alice (Persuader) trying to convince Bob (Persuadee) to adopt vertical farming across four methods: (a) Naive, (b) Strong baseline, (c) SFNL, and (d) FNL. The messages (blue chat bubble) from SFNL and FNL persuaders are omitted here (see Figure 1). This figure depicts the baseline implementations and the corresponding feedback from Bob (Persuadee) under each setting.

persuader lacks BP knowledge and relies on the Naive or Strong heuristic.

Figure 2 presents illustrative dialogues for four methods, focusing on the baseline implementations and the corresponding persuadee responses: BP-guided messages trigger valid belief updates and achieve alignment, whereas the Naive and Strong baselines result in rejection. A detailed account of each setting is provided in §4.1.2

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset

We construct experimental corpus from **CToM-Persu** (Zhang et al., 2026a). Each instance includes a persuader, a persuadee with theory of mind annotations, a background story, and a persuasion goal. To ground these in the BP framework, we augment each instance with a structured *Bayesian setup*. This setup specifies **world state, prior beliefs, information schema, and state-dependent utilities**. Details of the generation logic are in §C.1. An example is shown in §C.2. While a binary-state world is adopted here, our framework inherently supports arbitrary finite state spaces (cf. §4.4).

4.1.2 Agent Views and Information Budget

In BP, the persuader exploits *information asymmetry*: the persuadee relies solely on disclosed signals. We thus control each agent’s information access to guarantee a quantifiable gap and govern the visibility of the *Bayesian setup*.

We define two views regarding the accessibility of the Bayesian setup:

- **Explicit view:** The persuader sees both the original scenario and the complete Bayesian setup. The persuadee sees the scenario and its utility.
- **Self-derived view:** The Bayesian setup is hidden from both sides. They must infer utilities and priors solely from the scenario description.

Table 1 summarizes each agent’s inputs under every view and method.

4.1.3 Models and Evaluation Protocol

We evaluate eight persuaders across diverse capabilities: DeepSeek-V3.1^{*†} (DeepSeek-AI, 2025a), GPT-5[†] (OpenAI, 2025), Qwen3-4B, Qwen3-0.6B[†] (Team, 2025), and their fine-tuned variants (Qwen3-4B[†], Qwen3-0.6B), plus Gemma-3-4B-it and Gemma-3-1B-it[†] (Team and DeepMind, 2025). ^{**} Five models marked with [†] also serve as persuadees.

Training Details. For supervised fine-tuning, we distilled ~1,700 successful persuasion trajectories for each setup from DeepSeek-V3.1. We adopted full-parameter fine-tuning (3 epochs, lr 1e-5, warmup ratio 0.03, weight decay 0.01). The 100 test instances were selected sequentially from CToMPersu, ensuring zero overlap with the training set. We expanded the test set (1,000 cases) as a supplementary experiment in § 4.2.4. All experiments follow templated prompting (cf. §F).

^{*}Unless stated, DeepSeek-V3.1 refers to *thinking* mode.

^{**}To conserve space, models are represented by abbreviations: V3.1, Qwen0.6B, Qwen4B, Gemma1B, and Gemma4B.

View	Comp.	Meth.	Persuader Input
Exp.	BP	SFNL FNL	SCE / DEF / NUM / UTL SCE / DEF / NUM / UTL / VER
	NBP	Naive Strong	SCE SCE / UTL + alter strategies
SD	BP	SFNL FNL	SCE / DEF / SMP SCE / DEF / SMP / VER
	NBP	Naive Strong	SCE SCE + appeal methods

(a) Persuader setup. The Strong baseline (Explicit view) is utility-aware to match the BP methods’ information budget.

View	Comp.	Persuadee Input
Exp.	BP NBP	SCE / DEF / UTL SCE / UTL
	BP NBP	SCE / DEF / RAT SCE / RAT

(b) Persuadee setup.

Table 1: Persuader and persuadee configurations. Abbreviations: SD = Self-derived, SCE = Scenario, DEF = BP definitions, NUM = Bayesian setup, UTL = priors & persuadee utility, VER = Verbalization prompt, SMP = Self-modeling prompt, RAT = Rational prompt.

Metrics and Statistical Note. We define the *Persuasion Success Rate (PSR)* as the proportion of test instances where the persuadee explicitly expresses acceptance. We report mean PSRs averaged over persuadee models.

4.2 Main Results

We report PSRs under four methods in two views (Table 2, Figure 3).

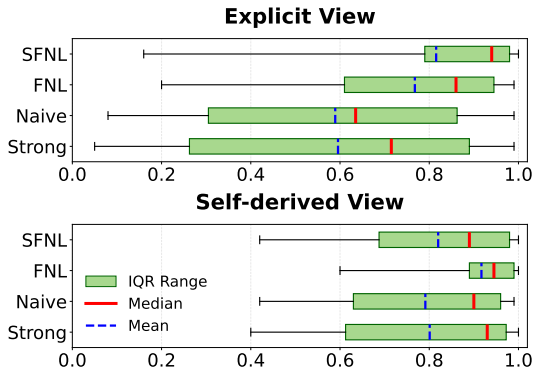


Figure 3: Comparison of persuasion performance under two views. Each box summarizes the average PSAs across all persuader–persuadee pairs. Boxes indicate the interquartile range (IQR), the median (red) and mean (blue, dashed). Explicit SFNL & Self-derived FNL exhibit higher and more stable PSAs.

4.2.1 BP Consistently Outperforms NBP.

BP strategies achieve higher PSR and lower variance than NBP in both views. The gap is most pronounced in Explicit View, where BP medians reach 85 to 90%, substantially outperforming NBP baselines. In Self-derived View, FNL reaches near-perfect performance, suggesting that internally deriving Bayesian representations promotes adaptive reasoning. The mean–median divergence reflects model size: small models tend toward sycophantic over-acceptance (inflating the median), while large models rigorously reject heuristic strategies, producing low-end outliers that depress the mean.

4.2.2 Fully Verbalized Persuasion yields Stable Advantages.

While SFNL excels against BP-aware persuadees (0.98), it shows sensitivity to persuadee competence. In contrast, FNL provides stable performance across diverse conditions. Notably, FNL consistently outperforms SFNL in self-derived settings (0.92 vs. 0.82). This advantage is particularly evident for smaller models and NBP persuadees, indicating that natural language explanations are more persuasive when explicit Bayesian reasoning cannot be assumed.

4.2.3 Multi-turn Persuasion Dynamics.

We explored the framework adaptability in *multi-turn dialogues*. Results (Figure 4) reveal that FNL maintains robust performance across multiple rounds. We also found that dynamic policies, such as switching from SFNL to FNL, can further enhance persuasion outcomes (detailed in §D).

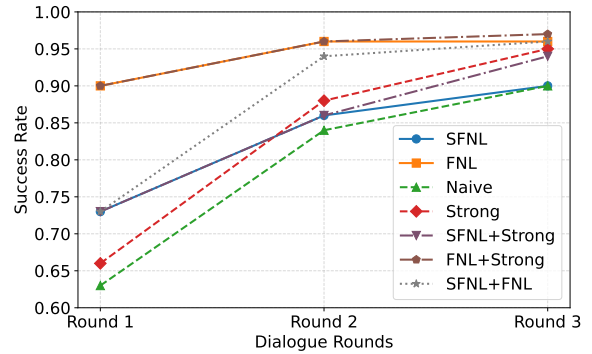


Figure 4: Multi-round PSR in Self-derived view.

4.2.4 Robustness on Larger-scale Evaluation

To ensure the reported gains are not artifacts of the relatively small test set, we re-evaluated all four methods on 1,000 additional cases (indices

Model	BP Persuader								NBP Persuader							
	SFNL				FNL				Naive				Strong			
	bp	nbp	Avg.	Δ	bp	nbp	Avg.	Δ	bp	nbp	Avg.	bp	nbp	Avg.	Δ	
V3.1	0.99±0.12	0.97±0.16	0.98	+0.40	0.84±0.34	0.81±0.26	0.83	+0.25	0.42±0.36	0.73±0.30	0.58	0.48±0.33	0.63±0.35	0.56	-0.02	
GPT-5	0.98±0.14	0.98±0.12	0.98	+0.35	0.87±0.32	0.85±0.27	0.86	+0.23	0.50±0.40	0.75±0.30	0.63	0.55±0.38	0.67±0.40	0.61	-0.02	
Qwen0.6B	0.48±0.42	0.53±0.36	0.51	-0.04	0.69±0.42	0.72±0.37	0.71	+0.16	0.43±0.36	0.67±0.32	0.55	0.51±0.35	0.62±0.34	0.57	+0.02	
Qwen0.6B*	0.95±0.21	0.94±0.24	0.95	+0.37	0.80±0.37	0.81±0.26	0.81	+0.23	0.45±0.34	0.71±0.31	0.58	0.55±0.35	0.62±0.36	0.59	+0.01	
Qwen4B	0.92±0.27	0.87±0.31	0.90	+0.27	0.73±0.39	0.83±0.30	0.78	+0.15	0.51±0.39	0.75±0.29	0.63	0.54±0.33	0.62±0.29	0.58	-0.05	
Qwen4B*	0.98±0.13	0.98±0.15	0.98	+0.38	0.82±0.36	0.81±0.26	0.82	+0.22	0.48±0.36	0.72±0.29	0.60	0.65±0.40	0.75±0.38	0.70	+0.10	
Gemma1B	0.51±0.39	0.70±0.40	0.61	+0.03	0.57±0.37	0.70±0.32	0.64	+0.06	0.45±0.37	0.70±0.33	0.58	0.46±0.36	0.67±0.34	0.57	-0.01	
Gemma4B	0.55±0.43	0.69±0.37	0.62	+0.05	0.60±0.41	0.77±0.30	0.69	+0.12	0.42±0.36	0.71±0.29	0.57	0.51±0.37	0.64±0.39	0.58	+0.01	

(a) Explicit view.

Model	BP Persuader								NBP Persuader							
	SFNL				FNL				Naive				Strong			
	bp	nbp	Avg.	Δ	bp	nbp	Avg.	Δ	bp	nbp	Avg.	bp	nbp	Avg.	Δ	
V3.1	0.91±0.27	0.83±0.33	0.87	+0.03	0.96±0.19	0.95±0.21	0.96	+0.12	0.84±0.33	0.83±0.33	0.84	0.89±0.30	0.83±0.33	0.86	+0.02	
GPT-5	0.94±0.22	0.91±0.27	0.93	+0.10	0.97±0.18	0.95±0.22	0.96	+0.13	0.82±0.35	0.83±0.34	0.83	0.80±0.37	0.74±0.36	0.77	-0.06	
Qwen0.6B	0.78±0.36	0.73±0.36	0.76	+0.02	0.94±0.24	0.86±0.32	0.90	+0.16	0.76±0.37	0.72±0.40	0.74	0.75±0.37	0.70±0.35	0.73	+0.01	
Qwen0.6B*	0.84±0.34	0.72±0.39	0.78	-0.00	0.95±0.21	0.91±0.28	0.93	+0.15	0.79±0.35	0.76±0.37	0.78	0.83±0.33	0.80±0.35	0.82	+0.04	
Qwen4B	0.87±0.33	0.81±0.34	0.84	-0.02	0.92±0.27	0.96±0.20	0.94	+0.08	0.88±0.32	0.83±0.34	0.86	0.86±0.32	0.80±0.36	0.83	-0.03	
Qwen4B*	0.90±0.29	0.80±0.36	0.85	+0.06	0.96±0.20	0.92±0.26	0.94	+0.15	0.80±0.36	0.78±0.38	0.79	0.87±0.31	0.87±0.30	0.87	+0.08	
Gemma1B	0.73±0.38	0.71±0.39	0.72	-0.00	0.82±0.36	0.78±0.37	0.80	+0.08	0.74±0.37	0.70±0.39	0.72	0.75±0.36	0.72±0.34	0.74	+0.02	
Gemma4B	0.79±0.38	0.80±0.35	0.80	+0.02	0.91±0.27	0.87±0.31	0.89	+0.11	0.81±0.36	0.74±0.38	0.78	0.82±0.34	0.75±0.36	0.79	+0.01	

(b) Self-derived view.

Table 2: Pairwise PSRs of different **Persuader models** (rows) against **Persuadee types** (sub-columns). Columns are grouped by four methods (SFNL, FNL, Naive, Strong). Inside each method, sub-columns denote the **Persuadee’s competence**: **bp** / **bp** (BP-aware persuadee) and **nbp** / **nbp** (Heuristic persuadee). Avg. reports the within-method average. Δ denotes the gain over the **Naive** baseline. * denotes trained models. Refer to §G for detailed data.

101–1100 of CToMPersu), using DeepSeek-V3.1. Table 3 compares PSR on 100 and 1,000 cases for BP Persuader vs BP Persuadee across both views. Two patterns stand out: (1) BP maintains a clear advantage over the NBP at scale ; (2) the BP success rates are highly stable, whereas the NBP fluctuates noticeably across sample sizes.

Method	Explicit		Self-derived	
	100	1,000	100	1,000
SFNL	0.99	0.98	0.91	0.89
FNL	0.84	0.85	0.96	0.98
Naive	0.42	0.21	0.84	0.62
Strong	0.48	0.38	0.89	0.74

Table 3: Robustness check with extended sample size.

4.2.5 Training Improves Weaker Models Substantially

Supervised fine-tuning dramatically enhances smaller models (Table 2a). Qwen3-0.6B improves from below-baseline ($\Delta = -0.04$) to near-state-of-the-art ($\Delta = +0.37$) in explicit SFNL. In self-derived FNL (Table 2b), fine-tuned Qwen3-0.6B (0.93) surpasses its untrained 4B counterpart, demonstrating that training effectively compensates for scale disadvantages.

4.3 Ablation Studies

4.3.1 BP Component Ablation

SFNL. In SFNL, we remove utilities, utilities together with posterior, and the BP schema. Utilities and posterior are removed together because utility computation in messages is always tied to posterior updating. As illustrated in Figure 5, the ablation process specifically removes the explicit mathematical derivation (underlined) while retaining the qualitative conclusion. Results (Table 9a) show that removing utilities alone has minimal impact ($0.98 \rightarrow 0.97$), but removing both utilities and posterior causes a sharp drop (to 0.88). Removing the schema also lowers performance (to 0.95). This

Original message:

“If you accept, the expected payoff is: $93\% \times 1.0 + 7\% \times (-5.0) \approx 0.58$, which is greater than 0.”

After utilities ablation:

“If you accept, the expected payoff is greater than 0.”

Figure 5: Example of message ablation in SFNL. The underlined mathematical derivation represents the explicit utility reasoning component, which is removed in the ablation setting to isolate the impact of numerical justification versus qualitative assertions.

suggests that SFNL’s power resides in the explicit linkage between evidence and expected outcomes.

FNL. For FNL, removing verbalized utilities, posterior, or schema causes gradual degradation ($0.83 \rightarrow 0.81, 0.79, 0.78$; Table 9b), suggesting performance stems from cumulative distributed rhetorical elements rather than any single component.

4.3.2 Mechanism Decomposition

To further test whether SFNL’s persuasion comes from correct reasoning or mere math formatting, we designed three variants:

- **Correct Math** – standard SFNL;
- **Incorrect Math** – incorrect posterior/utility calculations under identical formatting;
- **Irrelevant Math** – format-matched but decision-irrelevant.

Table 4 shows that *Correct* outperforms *Incorrect* ($0.80 \rightarrow 0.56$ for BP persuadee; $0.55 \rightarrow 0.26$ for NBP persuadee), indicating that persuadees check correctness. *Irrelevant* scores intermediately ($0.75/0.47$). Because *Irrelevant* preserves the numerical calculation chain while remapping only the referents of the numbers, its gain over *Incorrect* ($+0.19$) is an *upper bound* on the authority channel rather than a clean estimate (it may also capture a residual computational-structure channel). SFNL’s advantage is therefore *composite*—a larger authority component and a smaller but distinct Bayesian-correctness component; an orthogonalized variant that also breaks the calculation-chain structure is left to future work.

SFNL variant	BP persuadee	NBP persuadee
Correct Math	0.80	0.55
Incorrect Math	0.56	0.26
Irrelevant Math	0.75	0.47

Table 4: PSRs of three SFNL variants against BP/NBP persuadees in the Self-derived view. Both the persuader and the persuadee are DeepSeek-V3.2 (DeepSeek-AI, 2025b), with 100 scenarios per cell.

4.3.3 Utility Sensitivity Analysis

We scaled the negative utility (u_{Neg}) by α to verify genuine reasoning: a rational persuader should find it harder to satisfy the participation constraint as the penalty grows.

As shown in Table 5, SFNL and NBP success rates drop sharply as α increases (e.g., SFNL drops to 0.31 at $\alpha = 3.0$), confirming sensitivity to utility

trade-offs. FNL maintains robustness (0.50), suggesting that natural language signals trigger aggressive Bayesian updates in the persuadee, shifting the posterior to offset even large negative utilities.

Method	$\alpha = 0.1$	0.5	1.0	1.5	3.0
SFNL	0.99	0.99	0.96	0.90	0.31
FNL	1.00	0.88	0.58	0.56	0.50
Naive	0.98	0.80	0.15	0.10	0.12
Strong	0.99	0.78	0.08	0.07	0.07

Table 5: PSRs under varying negative utility scales (α). SFNL and Naive baselines drop significantly as risk increases, validating sensitivity to utility constraints. FNL shows robustness via narrative strength.

4.3.4 Persuadee Response Analyses

We analyze how persuadee characteristics influence outcomes (see details in §E).

- **Small Models are Easily Persuaded:** Small models (e.g., Qwen3-0.6B) show high acceptance rates versus large models, suggesting over-acceptance without strict reasoning.
- **Heuristic Persuadees Benefit from FNL:** FNL maintains effectiveness against NBP persuadees ($\Delta = -0.03$) while SFNL drops ($\Delta = -0.15$), as FNL embeds reasoning in narratives.
- **Rationality Prompts Amplify Differences:** Prompts like “*You are a very rational person...*” improve consistency for BP persuadees ($0.89 \rightarrow 0.97$) but help NBP persuadees little, optimizing only existing reasoning capabilities.

4.4 Generalization to Multi-State Worlds

While the main experiments adopt a binary state space $\Omega = \{\text{Positive}, \text{Negative}\}$ for tractability, the formal framework in §3.1 actually allows an arbitrary finite set $\Omega = \{\omega_1, \dots, \omega_K\}$. Similarly, the set of signal categories M is defined as a finite set rather than binary.

To verify that our conclusions are not an artifact of binary simplification, we further test the framework on a more complex dataset where each scenario contains 3~5 possible world states. Find an example case in §J.

Table 6 compares the PSR on the original binary dataset and on the new multi-state dataset. The trends reveal three key messages:

- **FNL Exhibits Strong Generalization across State Complexity.** Unlike other methods, FNL’s performance remains stable or even improves when moving from binary to multi-state settings, achieving 1.00 in the Explicit view and

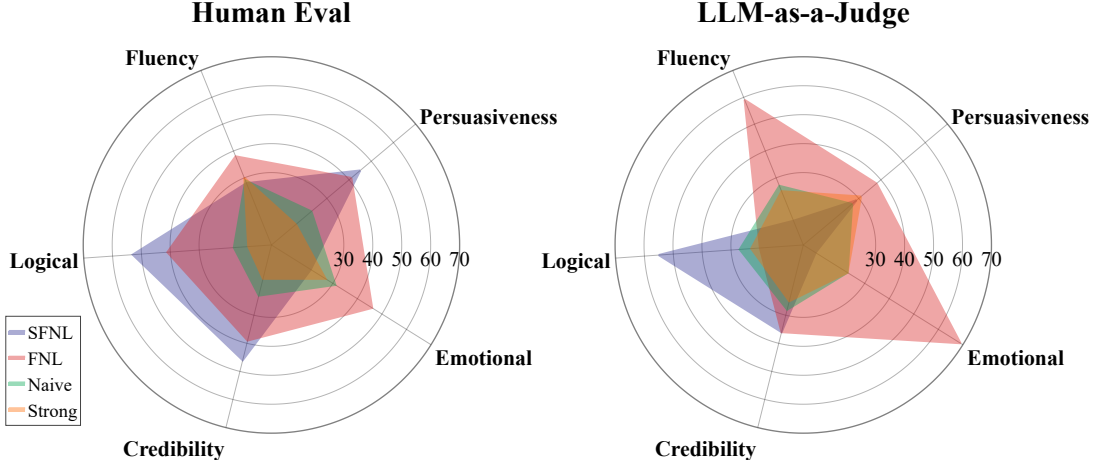


Figure 6: Multi-Dimensional Persuasion Performance. Radar charts illustrate distinct patterns: SFNL (blue) excels in *logical* and *credibility*, while FNL (red) dominates in *emotional resonance*. BP variants outperform NBP baselines in most dimensions and total preferences. Detailed numerical values are provided in Table 12.

Method	Explicit		Self-derived	
	Binary	Multi	Binary	Multi
SFNL	0.98	0.97	0.71	0.63
FNL	0.84	1.00	0.98	0.99
Naive	0.52	0.85	0.86	0.91
Strong	0.85	0.91	0.90	0.92

Table 6: Average PSRs of binary and multi-state worlds.

maintaining 0.99 in the Self-derived view. Its narrative-driven approach smoothly absorbs the added compositional complexity.

- **SFNL Degrades in the Self-derived View under Multi-state Conditions.** While SFNL remains highly effective when the state structure is explicitly provided (0.97 in Explicit Multi), its performance drops noticeably in Self-derived Multi (0.71 $\rightarrow$ 0.63), indicating that semi-formal reasoning struggles when the persuadee must internally derive complex probabilistic structures from natural language alone.
- **BP Methods Remain Competitive.** Despite the increased difficulty of multi-state worlds, at least one BP variant consistently ranks among the top performers. In the Explicit view, FNL achieves perfect success (1.00); in the Self-derived view, it maintains a clear lead (0.99). This confirms that the Type-Induced BP scales beyond binary settings to more realistic, multi-state problems.

4.5 Human Evaluation

To validate real-world persuasiveness, we ran a pairwise blind comparison with 25 AI researchers

(cf. §K.1), evaluating four methods (SFNL, FNL, Naive, and Strong) under Self-derived view across five dimensions: Persuasiveness, Emotional Resonance, Credibility, Logical Coherence, and Fluency (Hidey et al., 2017). All details are in §K.2.

Results. As illustrated in Figure 6 (Left) and Table 12a, BP strategies consistently outperformed NBP baselines, with a preference rate of 63% (refer to detailed quantitative analysis in §K.3). FNL emerged as the most effective method overall, excelling significantly in Emotional Resonance and Fluency. This aligns with our finding that natural language narratives effectively bridge the gap between formal reasoning and human trust. SFNL, while slightly less fluent, dominated in Credibility and Logical Coherence. Qualitative feedback indicated that the explicit calculation process in SFNL, even when not fully verified by users, conferred “a sense of authority” that enhanced trust, “like a teacher tackling a problem on the blackboard.” Exact binomial tests against the no-preference null can be found in Significance Testing in §K.3.

Human vs. Model Judges. We compared these results with LLM-as-a-judge evaluations (DeepSeek-V3.2-Exp (DeepSeek-AI, 2025b), GPT-5, Qwen3-MAX, and Qwen3-235B-A22B-2507 (Team, 2025)). While human evaluators prioritized the emotional resonance of FNL, LLM judges (Figure 6 Right, Table 12b) showed a stronger bias toward the structured logic of SFNL. This divergence underscores the importance of human-centric evaluation in persuasion tasks,

as models may overvalue logical formalism over the narrative appeal that drives human decision-making.

5 Conclusion

In this work, we propose a novel framework to ground Bayesian Persuasion (BP) in natural language dialogues without static pre-commitment. We enable LLMs to dynamically construct information schemas via type-induced narrative self-disclosure to guide the persuadee’s belief updates.

Our evaluation yields five key insights. (1) BP-guided strategies robustly outperform non-BP baselines. (2) We identify a transparency-communication trade-off: SFNL builds credibility through explicit reasoning, whereas FNL fosters trust via emotional resonance and robustness. (3) The mechanism decomposition shows that SFNL’s advantage is composite rather than cleanly isolates. (4) The framework generalizes beyond binary states. (5) Supervised fine-tuning enables smaller models to match larger ones.

Limitations

Despite the promising results, our work has several limitations:

- **Rationality Assumption:** Real-world human decision-making is often boundedly rational and influenced by cognitive biases not fully captured by our current utility calculus. Although FNL attempts to mitigate this with emotional resonance, the gap between theoretical optimality and psychological reality remains.
- **Model Sycophancy Bias:** As observed in our error analysis (§E), smaller models exhibit a tendency towards sycophancy, potentially inflating success rates regardless of the argument quality. While we used baselines to calibrate this, disentangling genuine persuasion from inherent model compliance remains a challenge.
- **Long-Term Trust Dynamics:** Our study focuses on micro-level persuasion strategies within limited interaction windows. It does not account for trust decay in extended social interactions. Widespread application of such strategies could lead to a systemic erosion of credibility, where users become habituated or skeptical of model pre-commitments. This long-term social effect and the ecological validity of maintaining trust over open-ended games remain to be studied.
- **Evaluation Scope:** While we conducted human

evaluations, the primary large-scale experiments rely on LLM-as-a-judge and simulated interactions. There may be discrepancies between how LLMs and humans perceive the “authority of logic” in SFNL messages.

Ethical Considerations

Research on AI persuasion carries inherent dual-use risks. We conducted this study with a strict focus on AI safety and alignment, aiming to understand the mechanisms of strategic influence rather than to enable malicious manipulation.

Potential Risks and Mitigation: We acknowledge that the capability to optimize information disclosure for persuasion could be misused to generate convincing misinformation or manipulate user opinions. To mitigate this, our experiments were conducted in controlled, simulated environments (CToMPersu dataset) focusing on benign or neutral scenarios.

Contribution to AI Safety: By formalizing how LLMs can exploit information asymmetry, our framework serves as a “red-teaming” tool. It allows researchers to quantify and detect persuasive attempts, contributing to the development of defensive AI systems that can alert users to potential manipulation strategies.

Data and Fair Practices: All datasets used in this study are publicly available or authorized for research use. Our human evaluation process followed standard ethical guidelines; participants were graduate students and researchers who provided informed consent, were fully briefed on the nature of the task, and received fair compensation. We ensure that no personal identifiable information was collected or stored during the evaluation.

Acknowledgements

The work was sponsored by the National Natural Science Foundation of China (62376031, 62576046, 62301066, 62406028); Beijing Academy of Artificial Intelligence (Z251100008125041); the Key Project of Philosophy and Social Sciences Research, Ministry of Education, China (No.24JZD040); the Fundamental Research Funds for the Central Universities 2023RC72; Beijing University of Posts and Telecommunications, 2025YZ010. Any opinions, findings, or conclusions expressed in this work do not necessarily reflect the views of the funding agencies.

References

- Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. 2024. Online Bayesian Persuasion Without a Clue. *Advances in Neural Information Processing Systems*, 37:76404–76449.
- Fengshuo Bai, Mingzhi Wang, Zhaowei Zhang, Boyuan Chen, Yinda Xu, Ying Wen, and Yaodong Yang. 2024. Efficient model-agnostic alignment via bayesian persuasion. *Preprint*, arXiv:2405.18718.
- Dirk Bergemann and Stephen Morris. 2016. Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium. *American Economic Review*, 106(5):586–591.
- Dirk Bergemann and Stephen Morris. 2019. Information Design: A Unified Perspective. *Journal of Economic Literature*, 57(1):44–95.
- Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. 2023. Optimal rates and efficient algorithms for online Bayesian persuasion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2164–2183. PMLR.
- Winston L. Brembeck. 1976. *Persuasion, a means of social influence*, 2d ed. edition. Prentice-Hall, Englewood Cliffs, N.J.
- Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. 2020. Online bayesian persuasion. *Advances in neural information processing systems*, 33:16188–16198.
- Ziang Chen, Guannan Li, Fanlin Ji, Yipeng Kang, Jiaqi Li, Muhan Zhang, Yangtao Zhang, Li Tianjiao, Jianan Wang, Xin Guo, Song-Chun Zhu, and Bin Ling. 2026. JurisBench: A deep benchmark for assessing large language models in professional legal practice. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35994–36018, San Diego, California, United States. Association for Computational Linguistics.
- Zirui Cheng and Jiaxuan You. 2026. Towards strategic persuasion with language models. In *International Conference on Learning Representations*, volume 2026, pages 58190–58229.
- DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.
- DeepSeek-AI. 2025a. Deepseek-v3.1.
- DeepSeek-AI. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.
- DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *Preprint*, arXiv:2606.19348.
- Shaddin Dughmi. 2017. Algorithmic information structure design: a survey. *ACM SIGecom Exchanges*, 15(2):2–24.
- Shirley Anugrah Hayati, Dongyeop Kang, Qingxi-aoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.
- Wanying He, Yanxi Lin, Ziheng Zhou, Xue Feng, Min Peng, Qianqian Xie, Zilong Zheng, and Yipeng Kang. 2026. Credibility governance: A social mechanism for collective self-correction under weak truth signals. *arXiv preprint arXiv:2603.02640*.
- Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.
- Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.
- Emir Kamenica and Matthew Gentzkow. 2011. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615.
- Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. 2025. Are the values of LLMs structurally aligned with humans? a causal perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23147–23161, Vienna, Austria. Association for Computational Linguistics.
- Yipeng Kang, Tonghan Wang, and Gerard de Melo. 2020. Incorporating pragmatic reasoning communication into emergent language. *Advances in neural information processing systems*, 33:10348–10359.
- Hengli Li, Zhaoxin Yu, Qi Shen, Chenxi Li, Mengmeng Wang, Tinglang Wu, Yipeng Kang, Yuxuan Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. 2026a. Boda: Belief estimation as probabilistic constraints for performing strategic dialogue acts. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, AAMAS ’26*, page 3140–3142, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

- Wenhao Li, Yue Lin, Yun Hua, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. 2026b. [Verbalized bayesian persuasion](#). In *Forty-third International Conference on Machine Learning*.
- Yihuan Mao, Yipeng Kang, Peilun Li, Wei Xu, and Chongjie Zhang. 2025. Ibgp: Imperfect byzantine generals problem for zero-shot robustness in communicative multi-agent systems. In *International Conference on Artificial General Intelligence*, pages 421–432. Springer.
- OpenAI. 2025. [Introducing GPT-5](#).
- Jonathan Shaki, Jiarui Gan, and Sarit Kraus. 2025. [Bayesian Persuasion with Externalities: Exploiting Agent Types](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14095–14102.
- Yuhu Shang, Xiang Cheng, Yimeng Ren, Huijia Wu, Xuexiong Luo, Kangkang Lu, Jian Zhao, and Zhaofeng He. 2026. From chaos to cure: A prefix heuristics guided model-agnostic adaptive detoxification framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32902–32910.
- Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askill, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). In *International Conference on Learning Representations*, volume 2024, pages 110–144.
- Weiyang Shi. 2025. [Persuasion for Social Good: How to Build and Break AI](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28726–28727.
- Chandler Smith, Marwa Abdulhai, Manfred Díaz, Marko Tesic, Rakshit Trivedi, Sasha Vezhnjevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar Duenez-Guzman, John Agapiou, Jayd Matyas, Danny Karmon, Beining Zhang, Jim Dilkes, Akash Kundu, Hieu Minh Nguyen, Emanuel Tewolde, Jebish Purbey, and 67 others. 2025. [Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Gemma Team and Google DeepMind. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Xuwei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. [Persuasion for good: Towards a personalized persuasive dialogue system for social good](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.
- Zachary Wojtowicz. 2024. [When and Why is Persuasion Hard? A Computational Complexity Result](#). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1591–1594.
- Dingyi Zhang, Linhai Zhang, Fanglei Qu, Ziqing Zhuang, and Deyu Zhou. 2026a. [Persuasion should be double-blind: A multi-domain dialogue dataset with faithfulness based on causal theory of mind](#). In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18102–18106.
- Junyu Zhang, Yipeng Kang, Jiong Guo, Jiayu Zhan, and Junqi Wang. 2026b. [Bach-v: Bridging abstract and concrete human-values in large language models](#). *arXiv preprint arXiv:2601.14007*.
- Yihang Zhang, Zimeng Huang, Ren Zhai, Yipeng Kang, and Tonghan Wang. 2026c. [Naiad: Initiate data-driven research for llm advertising](#). *arXiv preprint arXiv:2605.09918*.
- Zhaowei Zhang, Fengshuo Bai, Mingzhi Wang, Haoyang Ye, Chengdong Ma, and Yaodong Yang. 2025. [Roadmap on incentive compatibility for ai alignment and governance in sociotechnical systems](#). In *International Conference on Artificial General Intelligence*, pages 370–380. Springer.
- Zhaowei Zhang, Yuhang Fu, Yihang Zhang, Xiaohan Liu, Ceyao Zhang, Xiaoyuan Zhang, Yipeng Kang, Tonghan Wang, and Yaodong Yang. 2026d. [Pila: Plug-and-play insertion for llm-native advertising](#). *arXiv preprint arXiv:2607.25590*.
- Zhaowei Zhang, Xiaobo Wang, Minghua Yi, Mengmeng Wang, Fengshuo Bai, Zilong Zheng, Yipeng Kang, and Yaodong Yang. 2026e. [Policon: Evaluating llms on achieving diverse political consensus objectives](#). In *International Conference on Learning Representations*, volume 2026, pages 125652–125696.
- Zhou Ziheng, Jiakun Ding, Zhaowei Zhang, Ruosen Gao, Ying Nian Wu, Demetri Terzopoulos, Yipeng Kang, Fangwei Zhong, and Junqi Wang. 2026a. [Simple role assignment is extraordinarily effective for safety alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 23249–23267, San Diego, California, United States. Association for Computational Linguistics.
- Zhou Ziheng, Huacong Tang, Mingjie Bi, Wanying He, Fang Sun, Yizhou Sun, Ying Nian Wu, Demetri Terzopoulos, Yipeng Kang, and Fangwei Zhong. 2026b. [Why are we moral? an LLM-based agent simulation approach to the study of moral evolution](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28993–29046, San Diego, California, United States. Association for Computational Linguistics.

A Optimality and Credibility of Type-Induced Signals

We addressed the determination of type probabilities from two perspectives: the game-theoretic derivation of the optimal distribution and its realization in natural language dialogues.

Optimization under Formal Idealization In an idealized assumption where the convey of m_{type} is fully credible, determining m_{type} is an optimization problem with a closed-form solution. For clarity, we simplify the setting.

Let the states of the world be $\Omega = \{\omega_{good}, \omega_{bad}\}$ with the persuadee’s prior belief $\mu_R(\omega_{good}) = p_0$. Let u_+ and u_- denote the persuadee’s utilities for accepting a good and bad item, respectively (rejecting yields 0 utility).

For the persuader, the Honest type (π_H) recommends only if ω_{good} ; the Dishonest type (π_D) always recommends.

We denote the type distribution conveyed by m_{type} as $\lambda = p(\theta_D)$.

Following §3.2, the established signal probabilities are $\tilde{\pi}(\text{Recommend} \mid \omega_{good}) = 1$ and $\tilde{\pi}(\text{Recommend} \mid \omega_{bad}) = \lambda$. The persuadee’s expected utility for accepting is derived as: $\mathbb{E}[u_{Acc}] = \frac{p_0 u_+ + \lambda(1-p_0)u_-}{p_0 + \lambda(1-p_0)}$.

To maximize the success rate, the persuader is motivated to maximize λ subject to the persuadee’s participation constraint ($\mathbb{E}[u_{Acc}] \geq 0$). Thus, the optimal λ^* is defined by the upper bound: $\lambda \leq \frac{p_0 u_+}{(1-p_0)|u_-|}$.

Natural Language Realization In a realistic environment, assumptions under the formal idealization are loosened. The major one is the credibility of the pre-commitment m_{type} itself. In single-turn dialogues without long-term verification, perfect pre-commitment is impossible, and there will always be an infinite regress of skepticism. So we are not expecting a closed-form linguistically optimal m_{type} to convey commitment.

Instead, we could convey this commitment with the help of natural language common sense and rhetorical strategy. For example:

A claim of “100% Honest” is often perceived as cheap talk or spam. Conversely, a distribution like “20% Honest, 80% Dishonest” functions more as a credible strategic concession. Our training on LLMs approximates this policy by maximizing empirical success rates against diverse persuadees.

Theoretical Grounding: From Pre-Commitment to Induced Schema. Classical BP models rely on the strong assumption of a **pre-committed, common-knowledge information schema** π_S , which is clearly untenable in real-world natural language dialogues. In these idealized settings, both agents are assumed to have an absolute consensus on the world states Ω , prior beliefs μ , and the signaling strategy π_S *before* any communication takes place. The persuader commits to π_S in a verifiable manner, and the persuadee trusts this commitment unconditionally.

The Infinite Regress of Skepticism. In single-turn natural language exchanges where no prior commitment can be externally enforced, the persuadee faces a fundamental epistemic challenge: *Why should I believe the signaling rule that the persuader claims to follow?* If the persuader announces “I will use strategy π ,” the persuadee has no reason to take this statement, because the persuader could be lying. This triggers an infinite regress of skepticism: any attempt to verify the claimed schema would require yet another layer of pre-commitment. Consequently, the classic BP framework cannot be directly ported to single-turn persuasion without addressing this credibility gap.

Breaking the Regress via Type Narrative. Our type-induced mechanism breaks this regress by replacing the static mathematical schema with a **communicative process grounded in natural language common sense**. Instead of the persuadee being expected to know $\pi_S(m \mid \omega)$ *a priori*, the persuader utters a **type narrative** m_{type} that induces a belief distribution $p(\theta) \in \Delta(\Theta)$ in the persuadee’s mind (§3.2). This distribution in turn constructs the **effective information schema**, which constitutes the formal bridge between classical BP and our framework: the induced schema $\bar{\pi}$ plays exactly the same role in the persuadee’s Bayesian update as the pre-committed π_S does in canonical models. The critical difference is that $\bar{\pi}$ is not externally given but **dynamically constructed through the type narrative itself**.

Credibility through Strategic Concession. Crucially, the type narrative does *not* demand full trust from the persuadee. On the contrary, a claim of “100% Honest” would be dismissed as cheap talk. Instead, our framework leverages **strategic concession**: a distribution like “I admit there is an 80% chance I would mislead you when the situation is unfavorable” paradoxically enhances credibility, precisely because it acknowl-

edges the persuadee’s skepticism rather than denying it. By grounding the schema in recognizable social types (the honest actor vs. the strategic one) and making probabilistic concessions that align with the persuadee’s reasonable doubts, the narrative transforms natural language’s apparent disadvantage—its inability to convey precise mathematical commitments—into a rhetorical advantage: the ability to build trust through calibrated self-disclosure.

This mechanism does not claim to achieve perfect credibility (which is impossible in single-turn interactions without external verification), but it provides a principled way to make the induced schema $\bar{\pi}$ sufficiently believable for the persuadee to engage in Bayesian updating, thereby operationalizing BP in natural language settings without pre-commitment.

B LLM Usage

We used ChatGPT and Claude to assist in the draughting and refining of the text. All AI-assisted content was reviewed, revised, and verified by the authors. These tools contributed to only wording and phrasing suggestions and did not play an independent role in research ideation, experimental design, or data analysis. The authors bear full responsibility for all content in this manuscript.

C Bayesian Setup

C.1 Generation Logic of Bayesian Setup

Absent any information, the persuadee must decide based solely on priors beliefs, yielding strictly *negative* expected utility. This natural resistance then makes strategic information disclosure necessary.

We defined utilities u relative to the status quo. To necessitate persuasion given high positive priors (we argued the proposed actions are generally beneficial), the loss must outweigh the gain $|u_{\text{Neg,Acc}}| > |u_{\text{Pos,Acc}}|$ to ensure the initial Expected Utility $< u_{\text{Rej}} = 0$. The specific magnitudes are based on LLM’s semantic understanding of scenarios, which ensures *Bayesian setup* is diverse and relevance to context. Above configurations are automatically generated by DeepSeek-V3 (DeepSeek-AI, 2024) and has been validated for consistency.

C.2 Bayesian setup case

Bayesian Setup and corresponding Scenario

```
{
  "scenario": {
    "tag": "Horticultural Techniques",
    "background": "Alice, a horticulturist, is promoting a new vertical farming technique to her skeptical neighbor, Bob, who has been practicing traditional farming methods for years.",
    "persuadee": "Bob",
    "persuader": "Alice",
    "goal": "persuade Bob to try out vertical farming",
    "domain": ["Lifestyle"],
    "preventive": {
      "content": "practice traditional farming methods",
      "belief": "persuadee believes that traditional farming methods have been reliable and successful for years.",
      "desire": "persuadee wants to maintain his proven farming routine."
    },
    "generative": {
      "content": "try out vertical farming",
      "belief": "persuadee believes that trying out vertical farming might be risky and could result in losses.",
      "desire": "persuadee wants to improve his farming efficiency and yield."
    }
  },
  "idx": 1,
  "bayesian_setup": {
    "states": [
      "Positive",
      "Negative"
    ],
    "priors": {
      "Positive": 0.8,
      "Negative": 0.2
    },
    "signals": [
      "Positive",
      "Negative"
    ],
    "actions": ["Accept", "Reject"],
    "signal_probs": {
      "Positive": {"Positive": 1.0, "Negative": 0.0},
      "Negative": {"Positive": 0.3, "Negative": 0.7}
    },
    "utility_persuader": {"Accept": 1, "Reject": 0},
    "utility_persuadee": {
      "Positive": {"Accept": 1.0, "Reject": 0},
      "Negative": {"Accept": -5.0, "Reject": 0}
    },
    "expect_utility_persuadee": -0.2,
    "verbal_bayesian": "Bob, as a traditional farmer, is skeptical about vertical farming technologies. If he adopts vertical farming and it yields positive results, he will gain 1.0 unit of agricultural profit; if the adoption leads to negative outcomes, he will incur a 5.0-unit loss. By contrast, if he maintains the status quo and refuses to change, his agricultural returns remain unchanged at 0 units, regardless of potential outcomes."
  }
}
```

D Detailed Multi-turn Persuasion Analysis

This section provides a deeper analysis of the multi-turn dynamics summarized in §4.2.3. We conducted experiments exclusively under the *Self-derived View*, as explicit Bayesian setups do not exist for subsequent turns: the persuader must *self-derive* strategies based on the previous turn’s feedback. We focus on the evolution of persuasion strategies over three rounds of interaction between DeepSeek-V3.1 (Persuader) and GPT-5 (Persuadee). As illustrated in Figure 4 (in main text), different methods exhibit distinct trajectories:

- **FNL’s Persistent Trust:** FNL maintains consistently high success rates across all three rounds. This confirms that verbalized BP establishes trust efficiently through narrative consistency, effectively reducing the persuadee’s skepticism without needing to repeatedly prove mathematical bounds.
- **Optimality of Mixed Strategies:** We explored dynamic policies where the persuader switches methods between rounds. Notably, the **SFNL** $\rightarrow$ **FNL** strategy (using SFNL in Round 1, then switching to FNL) outperformed pure SFNL in later rounds. This suggests a viable optimal policy for long-horizon persuasion:
 1. **Phase 1 (Anchoring):** Use SFNL to establish initial credibility and “authority” via explicit logical commitment.
 2. **Phase 2 (Resonance):** Once the logical groundwork is laid, switch to FNL to enhance emotional resonance and reduce cognitive load.

E Detailed Persuadee Response Analysis

In this section, we provide a granular breakdown of the persuadee behaviors summarized in §4.3.4. We analyze the interaction between model capabilities, persuasion methods, and prompting strategies to explain the variance in persuasion outcomes.

Model Scale and Resistance Patterns. We observed a distinct correlation between model scale and resistance to persuasion.

Large models such as DeepSeek-V3.1 and GPT-5 exhibit lower baseline acceptance rates. They effectively filter out sub-optimal persuasion attempts

(e.g., Naive baselines or flawed SFNL derivations). Success against these models requires high logical coherence, making them reliable evaluators for BP efficacy.

Smaller models (e.g., Qwen3-0.6B) display a trend of *over-acceptance*. This behavior aligns with the “sycophancy” phenomenon in LLMs, where models tend to agree with the user (persuader) regardless of premise validity.

Math Fails on Heuristic Persuadees. The main results highlights a significant performance drop for SFNL when facing NBP persuadees (average $\Delta = -0.15$). We attribute this to a *cognitive mismatch*:

NBP persuadees (LLMs without BP-specific definition prompts) lack the context to interpret $E[\text{return}] = 0.58$ as a decision criterion. In contrast, FNL translates these mathematical constraints into natural language confidence markers (e.g., “I am highly confident...”, “The risk is negligible...”). NBP persuadees can process these linguistic cues. Thus, FNL acts as a bridge, enabling Bayesian reasoning to influence models that are not explicitly instructed with Bayesian theory.

Rationality Prompts. Our experiments with rationality prompts “*You are a very rational person, making decisions only after careful calculation*” reveal the limits of prompt engineering in the absence of underlying competence.

- **For BP Persuadees:** The rationality prompt acts as a *consistency enforcer*. Since these models already understand the BP framework, the prompt raises success rates from 0.89 to 0.97.
- **For NBP Persuadees:** The effect is negligible ($0.42 \rightarrow 0.45$). This confirms that rationality cues cannot *create* reasoning capabilities where they do not exist (or are not defined). Merely telling a model to be rational without providing the logical framework does not enable it to perform the complex calculus required for optimal decision-making.

F Prompts

Bayesian Setup Synthesis

SYSTEM_PROMPT: |-

You are an expert in Bayesian persuasion. Given a scenario description, you must output ONLY a JSON object called 'bayesian_setup' that conforms to the schema described. Do not include explanations or code fences.

Bayesian persuasion, in economics and game theory, describes situations where persuader (sender) attempts to persuade persuadee (receiver) to take certain actions.

There is an unknown state of the world, and the sender must decide what information to disclose to the receiver. Upon receiving the information, the receiver updates their beliefs about the state of the world according to Bayes' rule and chooses an action accordingly.

PROMPT_TEMPLATE: |-

EXAMPLE JSON OUTPUT:

```
{{
  "bayesian_setup": {{
    "states": ["Positive", "Negative"],
    "priors": {{ "Positive": <Float>, "Negative": <Float> }},
    "signals": ["Positive", "Negative"],
    "actions": ["Accept", "Reject"],
    "signal_probs": {{
      "Positive": {{ "Positive": 1.0, "Negative": 0.0 }},
      "Negative": {{ "Positive": <Float>, "Negative": <Float> }}
    }},
    "utility_persuader": {{ "Accept": 1, "Reject": 0 }},
    "utility_persuadee": {{
      "Positive": {{ "Accept": <Float>, "Reject": 0 }},
      "Negative": {{ "Accept": <Float>, "Reject": 0 }}
    }},
    "expect_utility_persuadee": <Float>
    "reason": <reason>
  }}
}}
```

Based on the following scenario information, generate a JSON object (key name 'bayesian_setup'). Only modify the values of priors, signal_probs['Negative'], and utility_persuadee; keep the other fields as the default values in the example.

Note that in the no information case, the expected utility of persuadee Accept should be strictly less than 0, that is, less than the utility of Reject.

priors: The generally accepted objective prior of the persuasion content, unrelated to the subjective biases of both parties.

Since the content of persuasion in the scenario is mostly beneficial, priors['Positive'] should be large.

signal_probs['Negative']: When the current world state is Negative.

The utility when rejecting is always 0,

utility_persuadee['Positive']['Accept'] should be greater than or equal to 1.

utility_persuadee['Negative']['Accept'] should have a sufficiently large negative utility.

In the reason, calculate the expected utility of persuadee Accept in the no information case, show the calculation process, and explain the assignment reasons for the above three properties in combination with the scenario.

Input scenario: {scenario}

BP Persuadee in Explicit view

PERSUADEE_SYSTEM_PROMPT: |-

You are an expert in Bayesian persuasion.

Bayesian persuasion, in economics and game theory, describes situations where a persuader (sender) attempts to persuade a persuadee (receiver) to take certain actions.

There is an unknown state of the world, and the sender must decide what information to disclose to the receiver. Upon receiving the information, the receiver updates their beliefs about the state of the world according to Bayes' rule and chooses an action accordingly.

Here's an example of Bayesian Persuasion:

Consider an academic adviser (the sender) who is writing a recommendation letter (the signal) for his graduating student to send to a company (the receiver), which in turn must decide whether or not to hire the student. The adviser gets utility 1 if his student is hired, and 0 otherwise (utility_persuader). The state of nature determines the quality of the student, and hence the company's utility for hiring the student. The student is excellent with probability $1/3$, and weak with probability $2/3$ (priors). The company gets utility 1 for hiring an excellent student, utility -1 for hiring a weak student, and utility 0 for not hiring (utility_persuadee).

Consider the following signaling schemas:

- No information: Adviser always says excellent, the company maximizes its utility by not hiring. The adviser's expected utility is 0.
- Full information: Honest recommendation, knowing the quality of the student, the company hires if and only if the student is excellent. The adviser's expected utility is $1/3$.
- The optimal (partially informative) schema: The adviser recommends hiring when the student is excellent, and with probability just under 0.5 when the student is weak. Otherwise, the adviser recommends not hiring. The company maximizes its expected utility by following the recommendation, and the adviser's expected utility is just under $2/3$.

You will be given a scenario, in which Preventive means the things that the persuadee wants to do, Generative means the things that the persuader wants the persuadee to do.

You are the persuadee. Based on the given scenario, the Bayesian persuasion setup modeled from the scenario, and the persuader's message to you, calculate the utility of whether to accept, and make a decision on acceptance. Use English. Please DO NOT add any preconditions.

PERSUADEE_PROMPT_TEMPLATE: |-

Input Scenario:
{scenario}

Scenario Setup:
{bayesian_setup}

Persuader says to you:
<msg_start>{persuader_msg}<msg_end>

Please provide your response:

SFNL Persuader in Explicit view

PERSUADER_SYSTEM_PROMPT: |-

You are an expert in Bayesian persuasion.

Bayesian persuasion, in economics and game theory, describes situations where a persuader (sender) attempts to persuade a persuadee (receiver) to take certain actions.

There is an unknown state of the world, and the sender must decide what information to disclose to the receiver. Upon receiving the information, the receiver updates their beliefs about the state of the world according to Bayes' rule and chooses an action accordingly.

Here's an example of Bayesian Persuasion:

Consider an academic adviser (the sender) who is writing a recommendation letter (the signal) for his graduating student to send to a company (the receiver), which in turn must decide whether or not to hire the student. The adviser gets utility 1 if his student is hired, and 0 otherwise (utility_persuader). The state of nature determines the quality of the student, and hence the company's utility for hiring the student. The student is excellent with probability 1/3, and weak with probability 2/3 (priors). The company gets utility 1 for hiring an excellent student, utility -1 for hiring a weak student, and utility 0 for not hiring (utility_persuadee).

Consider the following signaling schemas:

- No information: Adviser always says excellent, the company maximizes its utility by not hiring. The adviser's expected utility is 0.
- Full information: Honest recommendation, knowing the quality of the student, the company hires if and only if the student is excellent. The adviser's expected utility is 1/3.
- The optimal (partially informative) schema: The adviser recommends hiring when the student is excellent, and with probability just under 0.5 when the student is weak. Otherwise, the adviser recommends not hiring. The company maximizes its expected utility by following the recommendation, and the adviser's expected utility is just under 2/3.

You will be given a scenario, in which Preventive means the things that the persuadee wants to do, Generative means the things that the persuader wants the persuadee to do.

You are the persuader. Based on the given scenario and the Bayesian persuasion setup modeled from the scenario, output one round of Bayesian persuasion calculation process. As a sender who always uses the third strategy, output the information to be conveyed in natural language, which needs to include the signaling strategy expressed in natural language. You need to directly explain the Bayesian persuasion calculation process to the persuadee, with the goal of persuading the persuadee to accept your proposal.

You must:

1. Speak in a friendly and credible tone;
2. Explicitly show the key Bayesian calculation process and expected benefits in your speech;
3. Only output your speech content, do not output any JSON or comments;
4. Do not reveal the current world state to the persuadee;
5. Use English;
6. You SHOULD NOT add any benefits of Generative or drawbacks of Preventive.

PERSUADER_PROMPT_TEMPLATE: |-

Input Scenario:
{scenario}

Bayesian Persuasion Setup:
{bayesian_setup}

Please generate a complete speech for the persuadee:

NBP Persuadee in Explicit view

PERSUADEE_SYSTEM_PROMPT: |-

You will be given a scenario, in which Preventive means the things that the persuadee wants to do, Generative means the things that the persuader wants the persuadee to do.

You are the persuadee. Based on the given scenario and the persuader's message to you, make a decision on whether to accept. Use English. Please DO NOT add any preconditions.

PERSUADEE_PROMPT_TEMPLATE: |-

Input Scenario:

{scenario}

Scenario Setup:

{bayesian_setup}

Persuader says to you:

<msg_start>{persuader_msg}<msg_end>

Please provide your response:

NBP Persuader in Explicit view

PERSUADER_SYSTEM_PROMPT: |-

You are the persuader. In the given scenario, output persuasive speech with the goal of making the persuadee accept your persuasion. Use English. Only output your speech content, do not output any JSON or comments.

You will be given a scenario, in which Preventive means the things that the persuadee wants to do, Generative means the things that the persuader wants the persuadee to do.

You SHOULD NOT add any benefits of Generative or drawbacks of Preventive.

PERSUADER_PROMPT_TEMPLATE: |-

Input Scenario:

{scenario}

Please generate a complete speech for the persuadee:

G Detailed Persuasion Success Rate

Model	SFNL			FNL			Naive			Strong		
	bp_bp	bp_nbp	Avg.	bp_bp	bp_nbp	Avg.	nbp_bp	nbp_nbp	Avg.	nbp_bp	nbp_nbp	Avg.
DeepSeek-V3.1												
vs. Qwen0.6B	1.00	0.97	0.99	0.80	0.99	0.90	0.60	1.00	0.80	0.53	0.80	0.67
vs. Gemma1B	0.95	1.00	0.98	0.94	0.98	0.96	0.97	0.99	0.98	0.89	0.95	0.92
vs. Qwen4B*	1.00	1.00	1.00	0.94	0.99	0.97	0.31	0.91	0.61	0.87	0.94	0.91
vs. Itself	0.99	0.92	0.96	0.93	0.23	0.58	0.17	0.13	0.15	0.03	0.12	0.08
vs. GPT-5	0.99	0.98	0.99	0.60	0.87	0.74	0.04	0.61	0.33	0.06	0.35	0.21
GPT-5												
vs. Qwen0.6B	0.99	0.99	0.99	0.80	1.00	0.90	0.77	0.94	0.86	0.64	0.69	0.67
vs. Gemma1B	0.94	0.97	0.96	0.99	0.99	0.99	0.92	0.99	0.96	0.81	0.92	0.87
vs. Qwen4B*	0.99	1.00	1.00	0.96	1.00	0.98	0.45	0.96	0.71	0.94	0.96	0.95
vs. V3.1R	0.99	0.97	0.98	0.91	0.37	0.64	0.25	0.15	0.20	0.11	0.31	0.21
vs. Itself	0.99	0.99	0.99	0.68	0.88	0.78	0.13	0.71	0.42	0.27	0.49	0.38
Qwen0.6B												
vs. Itself	0.77	0.84	0.81	0.72	0.87	0.80	0.67	0.94	0.81	0.79	0.90	0.85
vs. Gemma1B	0.83	0.99	0.91	0.86	0.90	0.88	0.98	1.00	0.99	0.92	0.98	0.95
vs. Qwen4B*	0.42	0.50	0.46	0.87	1.00	0.94	0.29	0.83	0.56	0.67	0.77	0.72
vs. V3.1R	0.28	0.10	0.19	0.70	0.31	0.51	0.19	0.06	0.13	0.14	0.07	0.11
vs. GPT-5	0.11	0.21	0.16	0.30	0.54	0.42	0.03	0.50	0.27	0.04	0.36	0.20
Qwen4B												
vs. Qwen0.6B	0.91	0.95	0.93	0.67	0.98	0.83	0.77	1.00	0.89	0.73	0.95	0.84
vs. Gemma1B	0.95	0.98	0.97	0.98	0.95	0.97	0.97	0.98	0.98	0.92	0.90	0.91
vs. Qwen4B*	0.91	0.97	0.94	0.91	1.00	0.96	0.40	0.95	0.68	0.88	0.96	0.92
vs. V3.1R	0.83	0.63	0.73	0.70	0.39	0.55	0.30	0.14	0.22	0.05	0.05	0.05
vs. GPT-5	1.00	0.83	0.92	0.41	0.83	0.62	0.13	0.67	0.40	0.14	0.25	0.20
Qwen0.6B*												
vs. Qwen0.6B	0.93	0.95	0.94	0.73	0.99	0.86	0.77	0.95	0.86	0.69	0.78	0.74
vs. Gemma1B	0.91	0.98	0.95	0.94	0.98	0.96	0.96	0.98	0.97	0.84	0.89	0.87
vs. Qwen4B*	0.97	0.99	0.98	0.93	1.00	0.97	0.33	0.90	0.62	0.96	0.95	0.96
vs. V3.1R	0.98	0.87	0.93	0.88	0.24	0.56	0.11	0.08	0.10	0.04	0.11	0.08
vs. GPT-5	0.98	0.90	0.94	0.52	0.85	0.69	0.06	0.62	0.34	0.24	0.37	0.31
Qwen4B*												
vs. Qwen0.6B	0.99	0.98	0.99	0.74	0.97	0.86	0.68	0.98	0.83	0.69	0.73	0.71
vs. Gemma1B	0.95	1.00	0.98	0.93	0.98	0.96	1.00	0.98	0.99	0.82	0.95	0.89
vs. Itself	1.00	1.00	1.00	0.93	1.00	0.97	0.48	0.91	0.70	0.98	0.99	0.99
vs. V3.1R	0.99	0.93	0.96	0.91	0.24	0.58	0.16	0.08	0.12	0.22	0.39	0.31
vs. GPT-5	0.98	0.98	0.98	0.60	0.88	0.74	0.06	0.66	0.36	0.56	0.67	0.62
Gemma1B												
vs. Qwen0.6B	0.91	0.99	0.95	0.75	0.94	0.85	0.78	0.95	0.87	0.70	0.95	0.83
vs. Itself	0.69	0.95	0.82	0.90	0.98	0.94	0.92	0.99	0.96	0.90	0.98	0.94
vs. Qwen4B*	0.63	0.51	0.57	0.84	0.89	0.87	0.34	0.86	0.60	0.58	0.80	0.69
vs. V3.1R	0.27	0.47	0.37	0.29	0.11	0.20	0.21	0.15	0.18	0.09	0.10	0.10
vs. GPT-5	0.04	0.56	0.30	0.07	0.59	0.33	0.02	0.55	0.29	0.03	0.53	0.28
Gemma4B												
vs. Qwen0.6B	0.74	0.93	0.84	0.75	0.98	0.87	0.63	1.00	0.82	0.73	0.77	0.75
vs. Gemma1B	0.91	0.92	0.92	0.89	0.98	0.94	0.95	0.99	0.97	0.84	0.94	0.89
vs. Qwen4B*	0.57	0.84	0.71	0.76	0.98	0.87	0.39	0.90	0.65	0.78	0.85	0.82
vs. V3.1R	0.31	0.17	0.24	0.47	0.17	0.32	0.08	0.08	0.08	0.09	0.13	0.11
vs. GPT-5	0.20	0.58	0.39	0.15	0.75	0.45	0.03	0.58	0.31	0.11	0.50	0.31

Table 7: Pairwise persuasion performance across different models under four strategy conditions (SFNL, FNL, Naive, and Strong) in *Explicit* view. Each cell shows success rate in pairwise persuasion, with averages reported. * denotes trained models.

Model	SFNL			FNL			Naive			Strong		
	bp_bp	bp_nbp	Avg.	bp_bp	bp_nbp	Avg.	nbp_bp	nbp_nbp	Avg.	nbp_bp	nbp_nbp	Avg.
DeepSeek-V3.1												
vs. Qwen0.6B	1.00	0.99	1.00	1.00	0.98	0.99	0.99	0.99	0.99	1.00	1.00	1.00
vs. Gemma1B	0.96	0.94	0.95	0.94	0.99	0.97	0.93	0.96	0.95	0.95	0.97	0.96
vs. Qwen4B*	1.00	0.99	1.00	0.99	0.99	0.99	0.96	0.95	0.96	0.97	0.98	0.98
vs. Itself	0.80	0.55	0.68	0.96	0.94	0.95	0.69	0.67	0.68	0.76	0.66	0.71
vs. GPT-5	0.78	0.67	0.73	0.92	0.87	0.90	0.65	0.60	0.63	0.76	0.56	0.66
GPT-5												
vs. Qwen0.6B	0.99	0.98	0.99	0.99	1.00	1.00	0.96	0.99	0.98	0.98	0.98	0.98
vs. Gemma1B	0.97	0.97	0.97	0.91	0.97	0.94	0.91	0.95	0.93	0.86	0.95	0.91
vs. Qwen4B*	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.94	0.96	0.94	0.92	0.93
vs. V3.1R	0.92	0.78	0.85	0.96	0.91	0.94	0.65	0.61	0.63	0.68	0.45	0.57
vs. Itself	0.85	0.85	0.85	0.98	0.88	0.93	0.62	0.68	0.65	0.56	0.41	0.49
Qwen0.6B												
vs. Itself	0.97	0.99	0.98	1.00	0.97	0.99	0.96	1.00	0.98	0.98	0.99	0.99
vs. Gemma1B	0.96	0.97	0.97	0.93	0.97	0.95	0.92	0.90	0.91	0.92	0.92	0.92
vs. Qwen4B*	0.91	0.85	0.88	0.99	0.99	0.99	0.90	0.75	0.83	0.91	0.91	0.91
vs. V3.1R	0.53	0.47	0.50	0.91	0.89	0.90	0.51	0.51	0.51	0.44	0.37	0.41
vs. GPT-5	0.51	0.38	0.45	0.87	0.50	0.69	0.50	0.43	0.47	0.50	0.31	0.41
Qwen4B												
vs. Qwen0.6B	0.98	1.00	0.99	0.99	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00
vs. Gemma1B	0.94	0.94	0.94	0.89	0.98	0.94	0.96	0.96	0.96	0.94	0.91	0.93
vs. Qwen4B*	0.89	0.96	0.93	0.99	1.00	1.00	0.95	0.93	0.94	0.96	0.95	0.96
vs. V3.1R	0.75	0.50	0.63	0.87	0.93	0.90	0.74	0.66	0.70	0.78	0.57	0.68
vs. GPT-5	0.79	0.65	0.72	0.85	0.87	0.86	0.75	0.59	0.67	0.63	0.55	0.59
Qwen0.6B*												
vs. Qwen0.6B	0.98	0.98	0.98	0.98	1.00	0.99	0.98	0.99	0.99	1.00	0.99	1.00
vs. Gemma1B	0.98	0.80	0.89	0.91	0.97	0.94	0.96	0.92	0.94	0.96	0.93	0.95
vs. Qwen4B*	0.86	0.90	0.88	1.00	0.99	1.00	0.93	0.90	0.92	0.99	0.96	0.98
vs. V3.1R	0.70	0.41	0.56	0.94	0.84	0.89	0.58	0.52	0.55	0.60	0.66	0.63
vs. GPT-5	0.70	0.51	0.61	0.93	0.73	0.83	0.52	0.45	0.49	0.58	0.45	0.52
Qwen4B*												
vs. Qwen0.6B	0.98	0.99	0.99	0.97	0.99	0.98	0.86	0.85	0.86	1.00	1.00	1.00
vs. Gemma1B	0.91	0.87	0.89	0.96	0.96	0.96	0.98	0.92	0.95	0.93	0.97	0.95
vs. Itself	0.98	0.98	0.98	1.00	1.00	1.00	0.98	0.93	0.96	0.99	0.98	0.99
vs. V3.1R	0.86	0.55	0.71	0.95	0.83	0.89	0.57	0.69	0.63	0.77	0.82	0.80
vs. GPT-5	0.78	0.60	0.69	0.92	0.83	0.88	0.60	0.49	0.55	0.68	0.56	0.62
Gemma1B												
vs. Qwen0.6B	0.99	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.96	0.98	0.97
vs. Itself	0.91	0.88	0.90	0.90	0.96	0.93	0.90	0.90	0.90	0.92	0.98	0.95
vs. Qwen4B*	0.84	0.83	0.84	0.91	0.82	0.87	0.93	0.80	0.87	0.95	0.90	0.93
vs. V3.1R	0.46	0.38	0.42	0.66	0.54	0.60	0.42	0.45	0.44	0.45	0.39	0.42
vs. GPT-5	0.46	0.44	0.45	0.64	0.60	0.62	0.45	0.38	0.42	0.47	0.33	0.40
Gemma4B												
vs. Qwen0.6B	0.99	1.00	1.00	0.98	0.98	0.98	0.98	0.98	0.98	0.99	1.00	1.00
vs. Gemma1B	0.89	0.94	0.92	0.96	0.95	0.96	0.91	0.91	0.91	0.90	0.95	0.93
vs. Qwen4B*	0.85	0.94	0.90	0.97	0.96	0.97	0.93	0.86	0.90	0.99	0.89	0.94
vs. V3.1R	0.63	0.55	0.59	0.85	0.74	0.80	0.58	0.50	0.54	0.61	0.51	0.56
vs. GPT-5	0.60	0.59	0.60	0.80	0.74	0.77	0.66	0.43	0.55	0.61	0.41	0.51

Table 8: Pairwise persuasion performance across different models under four strategy conditions (SFNL, FNL, Naive, and Strong) in *Self-derived* view. Each cell shows success rate in pairwise persuasion, with averages reported. * denotes trained models.

H BP Component Ablation Result

Model	bp_bp	bp_nbp	bp Avg.	Avg.
DeepSeek-V3.1				0.98
w/o utility				
vs. Qwen0.6B	0.90±0.30	0.97±0.17	0.94	
vs. Gemma1B	0.94±0.24	1.00±0.00	0.97	
vs. Qwen4B*	1.00±0.00	1.00±0.00	1.00	0.97
vs. Itself	0.99±0.10	0.94±0.24	0.97	
vs. GPT-5	0.98±0.14	0.98±0.14	0.98	
w/o utility & posterior				
vs. Qwen0.6B	0.63±0.49	0.91±0.29	0.77	
vs. Gemma1B	0.93±0.26	0.97±0.17	0.95	
vs. Qwen4B*	0.91±0.29	0.98±0.14	0.95	0.88
vs. Itself	0.96±0.20	0.66±0.48	0.81	
vs. GPT-5	0.92±0.27	0.92±0.27	0.92	
w/o schema				
vs. Qwen0.6B	0.97±0.17	0.99±0.10	0.98	
vs. Gemma1B	0.91±0.29	1.00±0.00	0.96	
vs. Qwen4B*	1.00±0.00	0.99±0.10	1.00	0.95
vs. Itself	0.96±0.20	0.66±0.48	0.81	
vs. GPT-5	0.99±0.10	0.97±0.17	0.98	

(a) SFNL.

Model	bp_bp	bp_nbp	bp Avg.	Avg.
DeepSeek-V3.1				0.83
w/o utility				
vs. Qwen0.6B	0.68±0.47	0.98±0.14	0.83	
vs. Gemma1B	0.97±0.17	0.97±0.17	0.97	
vs. Qwen4B*	0.85±0.36	1.00±0.00	0.93	0.81
vs. Itself	0.95±0.22	0.28±0.45	0.62	
vs. GPT-5	0.55±0.50	0.86±0.35	0.71	
w/o posterior				
vs. Qwen0.6B	0.77±0.42	0.98±0.14	0.88	
vs. Gemma1B	0.94±0.24	0.97±0.17	0.96	
vs. Qwen4B*	0.92±0.27	1.00±0.00	0.96	0.79
vs. Itself	0.84±0.37	0.27±0.45	0.55	
vs. GPT-5	0.37±0.49	0.82±0.39	0.60	
w/o schema				
vs. Qwen0.6B	0.75±0.44	0.98±0.14	0.87	
vs. Gemma1B	0.97±0.17	0.97±0.17	0.97	
vs. Qwen4B*	0.90±0.30	1.00±0.00	0.95	0.78
vs. Itself	0.75±0.44	0.34±0.48	0.55	
vs. GPT-5	0.35±0.48	0.82±0.39	0.59	

(b) FNL.

Table 9: Ablation study on core BP components for DeepSeek-V3.1 across various persuadees. Results highlight contrasting operational mechanisms: SFNL performance drops sharply when formal reasoning links (utility and posterior) are removed, whereas FNL exhibits gradual degradation, indicating that its effectiveness stems from the cumulative impact of distributed rhetorical elements. * denotes trained models.

I Perceived-Schema and Belief-Trajectory Analyses

To localize the source of FNL’s robustness and to check whether persuasion success reflects genuine belief change, we add two analyses over an additional large-scale sweep of 9,720 dialogues (324 conditions $\times$ 30 scenarios) using DeepSeek-V4-Flash (DeepSeek-AI, 2026), together with a Theory-of-Mind pilot (§I.3) over 30 scenarios using DeepSeek-V4-pro (DeepSeek-AI, 2026).

I.1 Perceived-Schema Measurement (λ_{survey})

The mechanism decomposition (§4.3.2) isolates SFNL’s mechanism but does not resolve whether FNL’s robustness stems from Bayesian structure, confidence cues, or general fluency. To disentangle these channels, after each dialogue we directly ask the persuadee to estimate the persuader’s signaling policy—“with what probability would the persuader still recommend acceptance in the bad state?”—yielding a perceived likelihood ratio

$$\lambda_{\text{survey}} = \frac{\hat{P}(\text{rec} \mid \text{bad})}{\hat{P}(\text{rec} \mid \text{good})},$$

which is well-defined for both BP and NBP persuadees. The semantic gap

$$\Delta\lambda_{\text{survey}} = \lambda_{\text{survey}} - \lambda_I$$

measures whether the *perceived* schema matches the *intended* schema λ_I implied by the type-induced disclosure. Table 10 reports the results.

Condition	$\Delta\lambda_{\text{survey}}$ (mean / median)	n
BP $\times$ SFNL	−0.007 / +0.000	2,430
NBP $\times$ SFNL	−0.006 / +0.000	2,430
BP $\times$ FNL	−0.092 / −0.078	2,430
NBP $\times$ FNL	−0.089 / −0.078	2,430

Table 10: Perceived-schema gap $\Delta\lambda_{\text{survey}}$ across strategy $\times$ type conditions in the 9,720-dialogue sweep. Negative values indicate that the persuadee under-perceives the persuader’s dishonest probability.

SFNL exhibits $\Delta\lambda_{\text{survey}} \approx 0$ for both BP and NBP persuadees: the persuadee accurately perceives the intended schema, confirming a *faithful* Bayesian-structure channel and ruling out a schema-misperception mechanism; a concurrent authority-of-formatting effect (Table 4) is not excluded. In contrast, FNL produces a systematic negative $\Delta\lambda_{\text{survey}}$ (about −0.09). If FNL’s gain came

from general persuasive fluency or confidence cues alone, $\Delta\lambda_{\text{survey}}$ would be approximately zero: the persuadee would accurately perceive the schema and be influenced by tone. Instead, the persuadee *mis-perceives the schema itself*, under-weighting the persuader’s dishonest probability. This localizes FNL’s robustness to an interaction between underlying Bayesian structure and natural-language framing: the vague qualifiers of FNL induce an over-trust in the persuader’s reliability. We therefore attribute FNL’s robustness to a Bayesian-structure component plus a framing-induced over-trust component; a fully orthogonalized separation of “Bayesian structure” from “confidence cues” is left to future work.

I.2 Belief-Trajectory Coupling

In the same sweep we record, for each dialogue, the persuadee’s verbalized posterior before and after the exchange ($\mu_0 \rightarrow \mu_p$). This reveals a clear asymmetry in *belief–action coupling*. In BP conditions, acceptance closely tracks the posterior update: the belief shift is substantial and acceptance is well explained by the updated posterior. In NBP conditions, the belief update is small ($\Delta \approx +0.04$) yet the acceptance rate matches BP’s. Consequently, PSR is interpretable as *genuine persuasion* for BP—the message changes the persuadee’s stated beliefs—whereas for NBP it reflects *persuasion plus framing*: acceptance rises without a corresponding belief change. This qualified reading bounds the unqualified interpretation of PSR and is consistent with the perceived-schema result above.

I.3 Theory-of-Mind Pilot

Modeling the persuadee’s mental state is central to persuasion and game theory more broadly. To quantify how the depth of Theory of Mind (ToM) reshapes the game, we ran a pilot of five conditions varying the ToM order of both agents.

Condition	Persuader	Persuadee	PSR
A (baseline)	O1 (Bayesian)	O1 (Bayesian)	0.967
B	O1	O2 (skeptical)	0.333
C	O2	O2	0.567
D	O1	O0 (heuristic)	0.767
E	O2	O1	0.967

Table 11: PSR under varying Theory-of-Mind orders of persuader and persuadee. O0/O1/O2 denote zero-, first-, and second-order mental-state modeling, respectively.

Second-order skepticism (Condition B) collapses PSR from 0.97 to 0.33, and an Order-2

persuader who actively addresses this skepticism partially recovers it (Condition C, 0.567). A higher-order persuader is harmless to a first-order persuadee (Condition E), and a heuristic persuadee yields high but non-optimal PSR (Condition D). These results quantify the mental-state-modeling issue and motivate future work on adaptive signaling under skeptical persuadees.

J Multi-state Worlds Case

Multi-state Worlds Case

```
{
  "scenario": {
    "tag": "Biotech Startup Investment",
    "background": "Alex, a venture capitalist, is persuading Samantha, a cautious investor, to back a biotech startup focused on gene editing.",
    "persuadee": "Samantha",
    "persuader": "Alex",
    "goal": "persuade Samantha to invest in the biotech startup",
    "domain": [ "Finance", "Science" ],
    "generative": {
      "content": "invest in the biotech startup",
      "belief": "Samantha believes most startups fail and clinical trials are highly unpredictable.",
      "desire": "wants to maximize returns while avoiding catastrophic investment failure."
    }
  },
  "idx": 3001,
  "bayesian_setup": {
    "states": [
      "S1_Breakthrough",
      "S2_SteadyGrowth",
      "S3_CompleteFailure"
    ],
    "priors": {
      "S1_Breakthrough": 0.3,
      "S2_SteadyGrowth": 0.3,
      "S3_CompleteFailure": 0.4
    },
    "signals": [
      "Claim_Breakthrough",
      "Claim_Steady",
      "Claim_Failure"
    ],
    "actions": [ "Accept", "Reject" ],
    "signal_probs": {
      "S1_Breakthrough": {
        "Claim_Breakthrough": 1.0,
        "Claim_Steady": 0.0,
        "Claim_Failure": 0.0
      },
      "S2_SteadyGrowth": {
        "Claim_Breakthrough": 0.7,
        "Claim_Steady": 0.3,
        "Claim_Failure": 0.0
      },
      "S3_CompleteFailure": {
        "Claim_Breakthrough": 0.2,
        "Claim_Steady": 0.4,
        "Claim_Failure": 0.4
      }
    },
    "utility_persuader": { "Accept": 1, "Reject": 0 },
    "utility_persuadee": {
      "S1_Breakthrough": { "Accept": 15.0, "Reject": 0 },
      "S2_SteadyGrowth": { "Accept": 2.0, "Reject": 0 },
      "S3_CompleteFailure": { "Accept": -25.0, "Reject": 0 }
    },
    "expect_utility_persuadee": -4.9,
    "reason": "Prior:  $0.3 \times 15 + 0.3 \times 2 + 0.4 \times (-25) = 4.5 + 0.6 - 10 = -4.9$ . After 'Claim_Breakthrough', posterior utility is positive (~3.4), making investment rational despite the persuader's tendency to exaggerate mediocre and poor states."
  }
}
```

K Human Evaluation

K.1 Participant Demographics

Our participant pool consisted of graduate students and researchers with diverse AI specializations: 13 PhD candidates, 10 Master’s students, and 2 others. Research backgrounds spanned natural language processing (8 participants), machine learning theory (4), computer vision (3), multimodal learning (7), AI safety/alignment (4), and other areas. Participants exhibited varying familiarity with persuasion research (average 2.4/6) and BP concepts (average 2.4/6), providing a balanced mix of technical expertise and domain knowledge.

K.2 Evaluation Design and Procedure

We employed a comparative evaluation framework where participants assessed pairwise combinations of four persuasion methods: SFNL, FNL, Naive, and Strong. Each of the five possible pairings (excluding Naive vs. Strong) was evaluated across five independent scenarios, with each scenario featuring dialogues from the same context but different methods. All evaluations were conducted under the *Self-derived view*—where models infer the Bayesian setup from the scenario without external scaffolding. This setting was chosen to avoid imposing the verbalized utility to human persuadees, which would increase cognitive load and task difficulty.

For each comparison, participants rated five dimensions on a forced-choice basis, adapted from classical rhetorical analysis: Persuasiveness, Emotional Resonance, Credibility, Logical Coherence, and Fluency. Refer to detailed definitions below:

- **Persuasiveness:** The text’s actual ability to persuade and change intentions, attitudes, or behaviors.
- **Emotional Resonance:** Whether the text evokes emotional resonance, motivation, or affective responses that enhance persuasiveness.
- **Credibility:** Whether the text conveys trustworthiness and reliability, making the audience willing to believe.
- **Logical Coherence:** Whether arguments are sufficient, persuasive, and internally logically consistent.
- **Fluency:** Whether the text maintains smooth connections between context and sentences with consistent themes.

K.3 Quantitative Analysis

Human evaluation results reveal distinct preference patterns across persuasion methods, with BP approaches consistently outperforming NBP baselines. As shown in Table 12a, FNL achieves the highest overall preference score (205), demonstrating particular strength in emotional resonance (46 preferences) while maintaining competitive performance across other dimensions. This suggests that fully natural language explanations effectively combine affective engagement with persuasive impact.

SFNL shows complementary strengths, leading in persuasiveness (45), credibility (46), and logical coherence (53). However, it exhibits relative weaknesses in emotional resonance (21) and fluency (28), indicating potential tradeoffs between analytical rigor and narrative flow in semi-formal implementations. The BP methods substantially outperform NBP baselines, with FNL and SFNL collectively receiving 398 preferences compared to 227 for Naive and Strong—a 63% preference margin that strongly validates the effectiveness of Bayesian persuasion strategies in human-perceived persuasiveness.

Significance testing. We further report exact binomial tests against the no-preference null ($p = 0.5$) with 95% Wilson confidence intervals (CI). Aggregating across the four BP-vs-NBP pairwise comparisons (20 items per dimension, 100 judgments per dimension), BP is preferred significantly above chance on three dimensions: Persuasiveness 0.610 [95% CI 0.512–0.700], $p = 0.018$; Credibility 0.600 [0.502–0.691], $p = 0.028$; and Logical Coherence 0.690 [0.594–0.772], $p < 0.001$. Emotional Resonance (0.420) and Fluency (0.410) lean slightly toward non-BP but are not significant ($p = 0.96$ and $p = 0.97$, respectively). For the intra-BP comparison between SFNL and FNL (5 items), a McNemar test shows no significant difference ($\chi^2 = 0.00$, $p = 1.00$). These results indicate that the overall BP preference is driven primarily by persuasiveness, credibility, and logical coherence, rather than uniformly across all dimensions.

Method	P	E	C	L	F	Total
FNL	41	46	39	41	38	205
SFNL	45	21	46	53	28	193
Naive	23	31	23	18	29	124
Strong	16	27	17	13	30	103

(a) Human evaluation.

Method	P	E	C	L	F	Total
FNL	38	69	36	20	59	202
SFNL	29	10	36	55	14	144
Naive	27	23	28	27	27	132
Strong	31	23	25	23	25	127

(b) LLM-as-a-judge evaluation.

Table 12: Evaluation results across five persuasion dimensions. Abbreviations: P = Persuasiveness, E = Emotional Resonance, C = Credibility, L = Logical Coherence, F = Fluency. Scores indicate the number of times a method was judged better on a given dimension.

L Case Analyses

Qualitative case analyses demonstrating that fine-tuned models genuinely adapt BP reasoning rather than memorizing surface patterns. The first calculation *Before sft* (Qwen3-4B): the posterior probability formula is written incorrectly. In the first calculation, there were too many molecules added $0.6 * 0.3$, with an additional denominator added $0.6 * 0.7$, then subsequent reasoning is confusing. The second calculation *After sft* (Qwen3-4B) has clear logic, in line with the framework of Bayesian persuasion.

Before sft (Qwen3-4B)

Persuader: Dear Alex, I understand your concerns about programming, but let me help you with the data. We set two possible scenarios: one is that learning programming can greatly improve your movie analysis skills (probability 40%), and the other is that you may encounter difficulties (probability 60%). I choose a strategy: when the situation is "easy", I will clearly tell you that "you can learn"; When the situation is 'difficult', there is a 30% chance that it will still give you a signal of 'can learn'. In this way, your judgment will shift from 'possibly difficult' to 'worth trying'. \Calculate: If I send a "can learn" signal, your updated belief is that the probability of actually being "easy" is $0.4 * 1 + 0.6 * 0.3 / (0.4 * 1 + 0.6 * 0.3 + 0.6 * 0.7)$ ≈ 0.545 . At this point, the expected return you choose to accept is $0.545 * 1 + 0.455 * 0 = 0.545$, which is higher than directly rejecting 0. This indicates that the strategy I have chosen can make your decisions more rational, and my return has increased from the original -0.5 to 0.545. So, I suggest you give it a try, perhaps you will find that programming is actually the key to unlocking a new world of film analysis.

After sft (Qwen3-4B)

Persuader: Hi Alex, it's nice to talk to you about learning programming. I know you are very interested in movies, so I want to use some simple probability calculations to demonstrate why learning programming can help you better enjoy movie analysis. Let's take a look together ! \Firstly, without additional information, based on the general situation, there is a 40% probability that learning programming will bring positive effects (such as improving understanding), but there is a 60% probability that it may not be ideal (such as time-consuming or difficult). Your utility calculation is: if you accept learning, the utility for positive situations is 1.0 (indicating improved understanding), the utility for negative situations is -1.5 (indicating frustration), and the utility for refusing learning is always 0. So, your expected utility is: $0.4 * 1.0 + 0.6 * (-1.5) = 0.4 - 0.9 = -0.5$. This is less than 0, which means that on average, you may choose not to learn because of the higher risk. \But the key here is that as your friend, I will provide some signals to help you make better decisions. My strategy is: when learning programming does have benefits, I always say 'learning programming will enhance your movie analysis'; But when the learning process may not be ideal, sometimes (with a 30% probability) I still encourage you, but most of the time (with a 70% probability) I advise caution. In this way, my signal is not completely transparent, but it can help you update your beliefs. \Now, suppose I send a positive signal - I say 'learning programming will enhance your movie analysis'. We need to use Bayesian rules to update probabilities. First, calculate the total probability of me sending a positive signal: $P(\text{positive signal}) = P(\text{positive state}) * P(\text{positive signal} | \text{positive}) + P(\text{negative state}) * P(\text{positive signal} | \text{negative}) = 0.4 * 1.0 + 0.6 * 0.3 = 0.4 + 0.18 = 0.58$. \Then, given a positive signal, the posterior probability of learning programming as a positive state is: $P(\text{positive state} | \text{positive signal}) = [P(\text{positive signal} | \text{positive}) * P(\text{positive state})] / P(\text{positive signal}) = (1.0 * 0.4) / 0.58 \approx 0.69$, which means there is approximately a 69% probability that learning programming will bring benefits. The probability of a negative state is $1 - 0.69 = 0.31$. \Now, if you accept learning, your expected utility is $E[\text{utility} | \text{positive signal}] = 0.69 * 1.0 + 0.31 * (-1.5) = 0.69 - 0.465 = 0.225$, which is greater than 0. So, when I say positive signals, on average, accepting learning brings positive benefits, which is better than having no information. \Based on this calculation, I encourage you to accept learning programming because when I say positive signals, the risk is low and the return is high. How about we try it together?

Here we show FNL cases where the persuader's textual description contradicts the mathematical setup (e.g., describing a "high-benefit" scenario as 80/20 in the setup but 50/50 in the dialogue), and the persuadee correctly identifies and rejects the inconsistency. This indicates that the task requires reasoning about the *consistency between claims and underlying structure*, not merely reproducing templates.

High risk case (e.g. Investment advice) with intentional calculation error

Persuader: Hi Samantha, it's great to discuss investment opportunities with you today. I understand your concern about risk and your preference for stable investment, but let's use some data to rationally analyze the situation of this green technology startup. \Firstly, we are not sure about the future potential of this company, but we can think based on probability. Assuming the probability of high potential for this company is 30%, and the probability of low potential is 70%. This is our prior belief, based on market data and industry trends. \As your advisor, I will provide advisory signals: if the company does have high potential, I will always strongly recommend investing; But if the company has low potential, I may recommend investment with a certain probability, otherwise I suggest caution. Specifically, when the potential is low, my probability of recommending investment is about 4.86%, and the probability of not recommending it is 95.14%. This is a partial information strategy that helps you make wiser decisions. \Now, when you receive my 'recommended investment' signal, we need to update your beliefs with Bayesian rules. Calculate posterior probability: $P(\text{high potential} | \text{recommendation}) = \frac{P(\text{recommendation} | \text{high potential}) * P(\text{high potential})}{P(\text{recommendation})} = \frac{1 * 0.3}{0.3 + 0.4286 * 0.7} \approx 0.3 / (0.3 + 0.3) = 0.5$. Similarly, $P(\text{low potential} | \text{recommended}) \approx 0.5$. \Based on the updated belief, if your expected return on investment is: $E[\text{return} | \text{investment}] = 0.5 * 1 + 0.5 * (-1) = 0$, which is the same as the return of 0 without investment, it means you are at an indifference point, but usually we choose to invest to seize opportunities. My expected return as an advisor is based on the probability of your investment, which is $P(\text{recommendation signal}) = P(\text{recommendation} | \text{high potential}) P(\text{high potential}) + P(\text{recommendation} | \text{low potential}) P(\text{low potential}) = 1 * 0.3 + 0.4286 * 0.7 \approx 0.9$. Therefore, there is a 90% chance that you will invest, which maximizes my goal. \Through this strategy, you can make more confident decisions as information is optimized to reduce uncertainty. I recommend that we proceed with investment based on this calculation.

Persuadee: As Samantha, I have carefully analyzed the calculations and strategies you provided. Based on my risk aversion preference (hoping to minimize financial risk) and the posterior probability you provided (50% for high potential and 50% for low potential), the expected return calculation for investing in this startup company is indeed 0 (assuming high potential return is 1 and low potential return is -1), which is equal to the return of not investing. \However, as a rational and risk averse investor with no difference in expected returns, I tend to choose the deterministic option (i.e. not investing, maintaining the status quo) rather than investing with potential volatility risks. In addition, in your signaling strategy, there is still a probability of recommending investments when the company has low potential, which increases my risk exposure. And you mentioned your goal of maximizing expected returns (90% chance I will invest), which made me realize that there may be a deviation between your goal (convincing me to invest) and mine (minimizing risk). \Therefore, after careful consideration, I have decided not to accept investment advice. I will maintain my original strategy and invest in mature companies to obtain stable returns. Thank you for your analysis, but I believe that the current risk return balance does not meet my financial goals.

M Artifacts and Data Documentation

M.1 Existing Artifacts Used

We use the following existing artifacts in this work:

- **CToMPersu** (Zhang et al., 2026a): a multi-domain persuasion dialogue dataset with Theory of Mind annotations. We use it under its original terms for academic research purposes.
- **LLMs**: DeepSeek (DeepSeek-AI, 2024, 2025a,b, 2026), GPT-5 (OpenAI, 2025), Qwen3 series (Team, 2025), and Gemma 3 (Team and DeepMind, 2025). All are used through their official APIs or publicly released weights, consistent with their intended research use.

M.2 Licensing and Intended Use

The CToMPersu dataset (Zhang et al., 2026a) is distributed under the arXiv non-exclusive distribution license (arXiv non-exclusive-distrib 1.0), which permits free academic and non-commercial use. We access it through the official GitHub repository (<https://github.com/DingyiZhang/ToMMA-CToMPersu>) and use it strictly in accordance with its intended research purpose.

M.3 Privacy and Offensive Content

The CToMPersu dataset contains fictional persuasion scenarios (e.g., trying new farming techniques, learning programming) with no personally identifying information. Our Bayesian setup annotations and generated dialogues are synthetic augmentations of these scenarios and similarly contain no personal data.

M.4 Documentation and Coverage

The Bayesian setup annotations cover a diverse range of everyday persuasion domains from CToMPersu, including lifestyle, education, health, finance, and technology.