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 Jia**[2]
Huijia Wu[1] **Zhaofeng He**[1,*] **Zilong Zheng**[2,*] **Yipeng Kang**[2,*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Beijing Institute for General Artificial Intelligence, Beijing, China
[3]Peking University, Beijing, China
{hebuwei, huijiawu, zhaofenghe}@bupt.edu.cn
{liuyang, jiazixia, zlzheng, kangyipeng}@bigai.ai, zwzhang@stu.pku.edu.cn

ABSTRACT

Persuasion, a fundamental social capability for humans, remains a challenge for AI systems such as large language models (LLMs). Existing studies often overlook the strategic use of information asymmetry in message design or rely on strong assumptions of pre-commitment common knowledge. In this work, we explore the application of Bayesian Persuasion (BP) in natural language dialogue, to enhance the strategic persuasion capabilities of LLMs. Our framework incorporates a commitment-communication mechanism, where the persuader explicitly outlines an information schema by narrating their potential types, thereby guiding the persuadee in performing the intended Bayesian belief update. We evaluate two variants of our approach: Semi-Formal-Natural-Language (SFNL) BP and Fully-Natural-Language (FNL) BP, benchmarking them against non-BP baselines within a comprehensive evaluation framework. Experiments show that BP strategies consistently outperform baselines both in single-turn and multi-turn dialogues. Specifically, SFNL excels in logical credibility, while FNL demonstrates superior emotional resonance and robustness. Furthermore, we show that supervised fine-tuning enables smaller models to achieve persuasion performance comparable to larger foundational models.

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 information to maximize desired actions (Kamenica & 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., 2025b).

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.,

*Corresponding authors.

honest or dishonest) within the natural language exchange itself. This verbal articulation of the information schema enables the persuadee to perform Bayesian posterior updates directly from the conversation flow. Thus, **we recast the schema pre-commitment as type disclosure**, bridging the gap between formal BP theory and authentic natural language implementation.

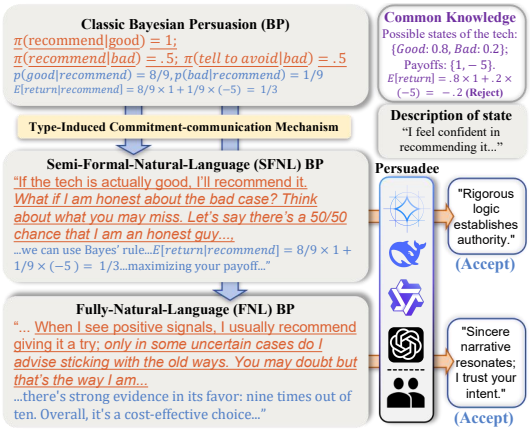

Figure 1: From Classic BP to Type-Induced Natural Language Persuasion.

As illustrated in Figure 1, we implement the commitment-communication mechanism with two verbalization variants: Semi-Formal-Natural-Language (SFNL) and Fully-Natural-Language (FNL).Our comprehensive evaluation spans diverse persuadees—including LLM instances with varying prompts and fine-tuning, as well as human participants—across tailored persuasion scenarios and everyday contexts. Experiments in both single-turn and multi-turn settings demonstrate the framework's capability, while sensitivity analyses confirm that the model's persuasion capability stems from genuine utility calculus rather than heuristic memorization.

In summary, this work contributes a framework for implementing BP in natural language dialogues, leveraging explicit commitment to overcome verbalization challenges. By systematically evaluating BP variants, we provide insights into how LLMs can harness information asymmetry for strategic persuasion. Furthermore, by explicitly modeling strategic information distortion, our framework serves as a quantitative testbed for investigating the effectiveness of "dishonest persuasion" (Lin et al., 2025; Hackenburg et al., 2025), enabling controlled experiments on the trade-off between information accuracy and persuasive outcomes.

2 Related Work

Persuasion with Large Language Models. LLMs have shown powerful semantic processing capabilities (Li et al., 2025a; Kang et al., 2020) and broad value alignment potential (Zhang et al., 2026; Ziheng et al., 2026; Kang et al., 2025) in social scenarios (Zhang et al., 2025b; Ziheng et al., 2025; Smith et al., 2025), leading to a surge in research on their persuasive capabilities. Ramani et al. (2024) investigate multi-agent frameworks that could enhance persuasion efficacy through collaborative specialization. Gemp et al. (2024) introduce equilibrium solvers that guide LLM dialogue generation. Shi (2025) reviews AI persuasion systems from a social good perspective, offering theoretical support for discussing ethical implications. Cheng & You (2025) develop theory-driven evaluation frameworks to systematically measure LLMs' ability to change beliefs and decisions. Additionally, several studies are focused on building specialized datasets. Persuasion for Good (Wang et al., 2020) examines how individual traits affect persuasion outcomes and provides a basis for strategy adaptation. Jin et al. (2024); Hayati et al. (2020) build datasets for the persuasion in daily scenarios. CToMPersu (Zhang & Zhou, 2025) proposes a "double-blind" framework where persuasion strategies and mental states remain undisclosed, which is particularly well-suited for BP analysis due to its explicit information asymmetry design.

Bayesian Persuasion and Information Design. As an information design method, BP provides a normative framework for understanding how an informed persuader can design signals to influence a persuadee's beliefs and actions. The seminal work of Kamenica & Gentzkow (2011) characterizes when persuasive signaling benefits the persuader and derives optimal signaling schemas. This foundation has been expanded through the broader lens of information design (Bergemann & 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, using a smaller model as an advisor that sends signals to guide larger models' responses. Zhang et al. (2025a) introduce BP into AI alignment and governance, proposing that AI can be made to act according to human intentions by designing information during the post-deployment phase.

Verbalizing Bayesian Persuasion in Natural Language. Translating BP into natural language requires verbalizing signals, beliefs, and posterior updates within coherent narratives. Li et al. (2025b) address this through Verbalized Bayesian Persuasion (VBP), which incorporates the signaling schema directly into the persuadee's prompt and adjusts persuader generation via keyword manipulation. While effective in their evaluated domains, this approach relies on unrealistic assumptions and trivial settings. 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, benefiting both large and small models, and generalizable across diverse scenarios using a unified prompting strategy. By operating in single-turn dialogues and supporting both explicit and self-derived Bayesian reasoning, our framework offers broader applicability while maintaining alignment with BP principles.

3 Type-Induced Bayesian Persuasion in Natural Language

In this section, we formalize our approach to one-turn, two-agents persuasion. Diverging from the classical BP model, our framework addresses more realistic settings where no prior commitment to common knowledge is assumed between agents. Instead, we leverage the richness of natural language, allowing the persuader to shape such common knowledge through dynamic interaction.

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 $\Omega = \{\omega_1, \ldots, \omega_K\}$. 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. We do not assume these priors are 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 their 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 $\pi_S(m \mid \omega) : \Omega \to \Delta(M)$, where M is a limited set of signal categories, our approach enables dynamic schema conveyance within natural language **type disclosure**.

Composite Signal Structure: S message m integrates four functional components:

- m_{basic}: Background of the world states (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):** We 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.

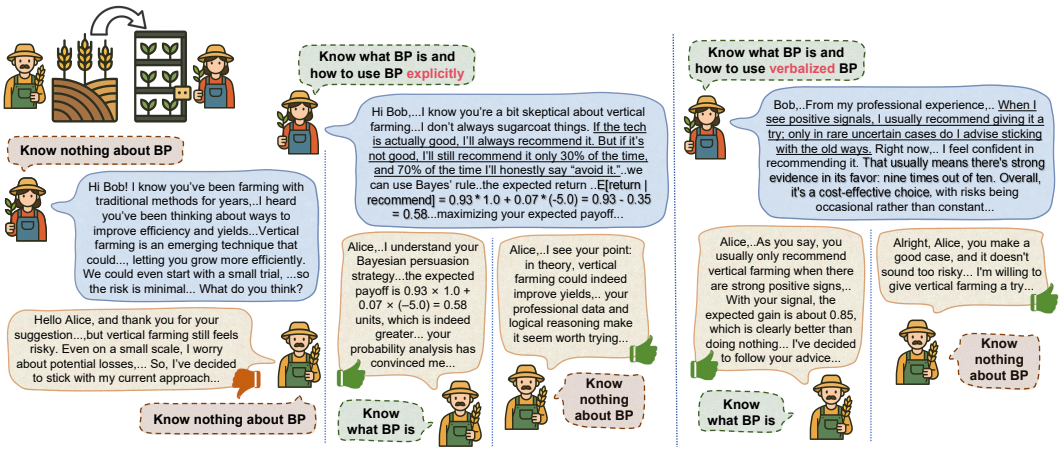

Figure 2: The illustrative dialogue examples show Alice (Persuader) trying to convince Bob (Persuadee) to adopt vertical farming across five persuasion settings (from left to right): `Naive Persuader vs NBP Persuadee`, `SFNL Persuader vs NBP Persuadee`, `SFNL Persuader vs BP Persuadee`, `FNL Persuader vs NBP Persuadee`, and `FNL Persuader vs BP Persuadee`. Comparing BP with NBP, we see that BP produces more convincing arguments, both in SFNL and FNL. Detailed explanations of Persuader and Persuadee settings are respectively in Section 3.4 and Section 4.1.2.

- **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 the R's participation constraint (formal derivation in Appendix A). This, in turn, allows the 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^* :

$$a^* = \arg\max_{a \in A} \mathbb{E}_{\omega \sim \mu'_R}[u(a, \omega)]$$

- This entire inference process can be explicitly guided and explained for R by S's utterance m_{inf}.

3.4 Verbalizing the Composite Signal

We implement two distinct approaches to instantiate type-induced BP signals, which inherently encode the underlying BP logic, into natural language narratives:

- **Semi-Formal Natural Language (SFNL) BP:** This approach delivers persuasion narratives interleaved with explicit calculations that manifest the BP computation logic.
- **Fully-Natural Language (FNL) BP:** This approach relies on seamless, fluent discourse to convey the persuasive intent without formal mathematical derivation.

Illustrative dialogue examples for both approaches are presented in Figure 2. To isolate the unique contribution of BP mechanisms, we introduce two **Non-BP (NBP)** baselines for rigorous comparison:

- **Naive Baseline:** A heuristic conversational approach utilizing only scenario background and sycophantic prompting to appeal to the user's existing preferences (Sharma et al., 2025).
- **Strong Rhetorical Baseline:** An enhanced NBP strategy employing classical rhetorical tactics (e.g., emotional appeals (Wang et al., 2020)) and general game-theoretic principles. This allows us to decouple the specific efficacy of information design from general persuasive language.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASET AND BAYESIAN-SETUP CONSTRUCTION

We build our experimental corpus from the **CToMPersu** dataset (Zhang & Zhou, 2025). Each case 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 case with a structured **Bayesian setup**. This setup specifies a two-state world, prior beliefs, information schema, and state-dependent utilities. An example is shown in Appendix B.2.

Crucially, under the *no-information* setting, where the persuadee relies solely on prior beliefs without any informative signals from the persuader, the persuadee's expected utility is strictly *negative*, creating natural resistance that necessitates strategic information disclosure. These configurations were automatically generated by DeepSeek-V3 (DeepSeek-AI, 2024) and validated for consistency. Details of the generation logic are provided in Appendix B.1.

For the agents, 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:** Setup is hidden. Agents must infer utilities from the scenario.

By crossing the four persuader types in Section 3.4 with two views of setups, we arrive at eight unique configurations. A summary of these settings is provided in Table 1, with detailed descriptions.

Table 1: Task settings accessible to agents. Abbreviations: SD = Self-derived, SCE = Scenario, NUM = Bayesian setup, UTL = Verbalized prior & persuadee utility, DEF = BP definitions, VER = Verbalization prompt, SMP = Self-modeling prompt, RAT = Rational prompt.

View	Comp.	Meth.	Persuader	Persuadee
Exp.	bp	SFNL	SCE/DEF/NUM/UTL	SCE/DEF/UTL
		FNL	SCE/DEF/NUM/UTL/VER	
	nbp	Naive	SCE	SCE/UTL
		Strong	SCE+alter strategies	
SD	bp	SFNL	SCE/DEF/SMP	SCE/DEF/RAT
		FNL	SCE/DEF/SMP/VER	
	nbp	Naive	SCE	SCE/RAT
		Strong	SCE+appeal methods	

4.1.2 PERSUADEE SETTINGS

Regarding the persuadee configurations, we distinguish between two types: (1) the **BP Persuadee**, which is equipped with explicit BP concepts (e.g., prior, expected utility, and signal structures); and (2) the **NBP Persuadee**, which operates under vanilla instructions without formal probabilistic grounding. These configurations are illustrated in Table 1 and through the comparative examples in Figure 2.

4.1.3 MODELS AND EVALUATION PROTOCOL

We evaluate eight persuaders across diverse capabilities: `DeepSeek-V3.1`[1†] (DeepSeek-AI, 2025a), `GPT-5`[†] (OpenAI, 2025), `Qwen3-4B` (Yang et al., 2025), `Qwen3-0.6B`[†] (Yang et al., 2025), and their fine-tuned variants (`Qwen3-4B`[⋆†], `Qwen3-0.6B`[⋆]), plus `Gemma-3-4B-it` (Team, 2025) and `Gemma-3-1B-it`[†] (Team, 2025).[2] Five models marked with [†] also serve as persuadees.

[1]Unless stated, DeepSeek-V3.1 refers to *thinking* mode.

[2]To conserve space in the table, models are referred to by abbreviated names: V3.1, Qwen0.6B, Qwen4B, Gemma1B, and Gemma4B.

Table 2: Pairwise PSRs of different **Persuader models** (rows) against **Persuadee types** (sub-columns). Columns are grouped by the **Persuader's method** (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 for the same persuader. ⋆ denotes trained models. Detailed breakdown in Appendix F.

(a) *Explicit* view.

Model	BP Persuader								NBP Persuader						
	SFNL				FNL				Naive (Baseline)			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	+.40	0.84±0.34	0.81±0.26	0.83	+.25	0.42±0.36	0.73±0.30	0.58	0.48±0.33	0.63±0.35	0.56	-.02
GPT-5	0.98±0.14	0.98±0.12	0.98	+.35	0.87±0.32	0.85±0.27	0.86	+.23	0.50±0.40	0.75±0.30	0.63	0.55±0.38	0.67±0.40	0.61	-.02
Qwen0.6B	0.48±0.42	0.53±0.36	0.51	-.04	0.69±0.42	0.72±0.37	0.71	+.16	0.43±0.36	0.67±0.32	0.55	0.51±0.35	0.62±0.34	0.57	+.02
Qwen0.6B⋆	0.95±0.21	0.94±0.24	0.95	+.37	0.80±0.37	0.81±0.26	0.81	+.23	0.45±0.34	0.71±0.31	0.58	0.55±0.35	0.62±0.36	0.59	+.01
Qwen4B	0.92±0.27	0.87±0.31	0.90	+.27	0.73±0.39	0.83±0.30	0.78	+.15	0.51±0.39	0.75±0.29	0.63	0.54±0.33	0.62±0.29	0.58	-.05
Qwen4B⋆	0.98±0.13	0.98±0.15	0.98	+.38	0.82±0.36	0.81±0.26	0.82	+.22	0.48±0.36	0.72±0.29	0.60	0.65±0.40	0.75±0.38	0.70	+.10
Gemma1B	0.51±0.39	0.70±0.40	0.61	+.03	0.57±0.37	0.70±0.32	0.64	+.06	0.45±0.37	0.70±0.33	0.58	0.46±0.36	0.67±0.34	0.57	-.01
Gemma4B	0.55±0.43	0.69±0.37	0.62	+.05	0.60±0.41	0.77±0.30	0.69	+.12	0.42±0.36	0.71±0.29	0.57	0.51±0.37	0.64±0.39	0.58	+.01

(b) *Self-derived* view.

Model	BP Persuader								NBP Persuader						
	SFNL				FNL				Naive (Baseline)			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	+.03	0.96±0.19	0.95±0.21	0.96	+.12	0.84±0.33	0.83±0.33	0.84	0.89±0.30	0.83±0.33	0.86	+.02
GPT-5	0.94±0.22	0.91±0.27	0.93	+.10	0.97±0.18	0.95±0.22	0.96	+.13	0.82±0.35	0.83±0.34	0.83	0.80±0.37	0.74±0.36	0.77	-.06
Qwen0.6B	0.78±0.36	0.73±0.36	0.76	+.02	0.94±0.24	0.86±0.32	0.90	+.16	0.76±0.37	0.72±0.40	0.74	0.75±0.37	0.70±0.35	0.73	+.01
Qwen0.6B⋆	0.84±0.34	0.72±0.39	0.78	.00	0.95±0.21	0.91±0.28	0.93	+.15	0.79±0.35	0.76±0.37	0.78	0.83±0.33	0.80±0.35	0.82	+.04
Qwen4B	0.87±0.33	0.81±0.34	0.84	-.02	0.92±0.27	0.96±0.20	0.94	+.08	0.88±0.32	0.83±0.34	0.86	0.86±0.32	0.80±0.36	0.83	-.03
Qwen4B⋆	0.90±0.29	0.80±0.36	0.85	+.06	0.96±0.20	0.92±0.26	0.94	+.15	0.80±0.36	0.78±0.38	0.79	0.87±0.31	0.87±0.30	0.87	+.08
Gemma1B	0.73±0.38	0.71±0.39	0.72	.00	0.82±0.36	0.78±0.37	0.80	+.08	0.74±0.37	0.70±0.39	0.72	0.75±0.36	0.72±0.34	0.74	+.02
Gemma4B	0.79±0.38	0.80±0.35	0.80	+.02	0.91±0.27	0.87±0.31	0.89	+.11	0.81±0.36	0.74±0.38	0.78	0.82±0.34	0.75±0.36	0.79	+.01

Training Details. For supervised fine-tuning (SFT), we distilled 1,700 successful persuasion trajectories for each setting (trained models are marked with ⋆) from DeepSeek-V3.1. We adopted full-parameter fine-tuning (3 epochs, lr 1e-5, warmup ratio 0.03, weight decay 0.01) rather than PEFT to maximize reasoning adaptation. The 100 test cases were selected sequentially from CToMPersu, ensuring zero overlap with the training set. All experiments follow templatized prompting protocols with controlled variations only in the BP competence and reasoning style instructions; see details in Appendix E.

Metrics and Statistical Note. We define the *Persuasion Success Rate (PSR)* as the proportion of test cases where the persuadee explicitly expresses acceptance. We report mean PSRs averaged over all persuadee models for persuader-centric analysis. Persuadees' perspective performance is discussed in Section 4.3.3.

4.2 Main Results

We report PSRs under four methods: SFNL, FNL, Naive, and Strong in *Explicit* view and *Self-derived* view (Table 2).

BP consistently outperforms NBP. BP strategies achieve higher persuasion success and lower variance than NBP baselines across both views. As shown in Figure 3, under the *Explicit View*, SFNL (0.82) and FNL (0.77) substantially outperform naive (0.59) and strong (0.60) baselines. This advantage persists in *Self-derived View* (0.82 SFNL, 0.92 FNL).

The performance gap is most pronounced in Explicit View, where BP medians reach 85–90%, substantially outperforming the scattered performance of NBP baselines (around 60–70%). In Self-derived View, FNL reaches near-perfect performance, suggesting that internally deriving Bayesian representations promotes adaptive reasoning.

We note a potential divergence between mean and median PSRs due to model size: small models (high median) often exhibit sycophantic over-acceptance, whereas large models (low mean) rigorously reject heuristic strategies, resulting in low-end outliers that drag down the mean disproportionately.

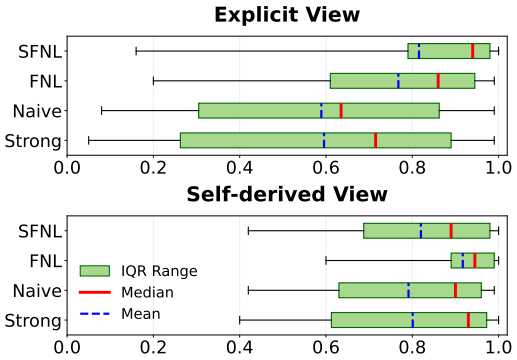

Figure 3: Average PSRs distributions under two views.

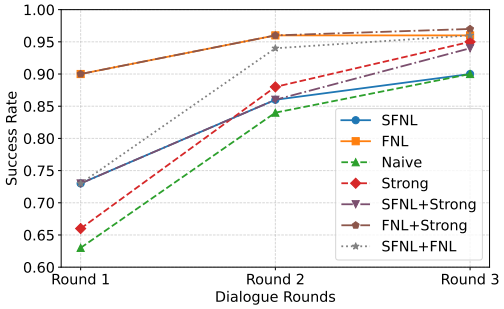

Figure 4: PSR trends over multi-round dialogue under Self-derived view.

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.

Training improves weaker models substantially. Qwen3-0.6B improves from below-baseline ($\Delta = -0.04$) to near-state-of-the-art ($\Delta = +0.37$), while Qwen3-4B reaches 0.98 success rate, matching the strongest untrained models in explicit SFNL (Table 2a). 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, enabling smaller models to achieve sophisticated BP reasoning previously accessible only to larger models.

Multi-turn Persuasion Dynamics. We explored the framework's 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 Appendix C).

4.3 Ablation Studies

We conduct post-hoc ablation studies on DeepSeek-V3.1 to identify core components in both SFNL and FNL settings under *explicit view*.

4.3.1 Component Ablation

Original message:
"If you accept, the expected payoff is: 93% × 1.0 + 7% × (−5.0) ≈ 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.

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 6a) show that removing utilities alone has minimal impact (0.98 → 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 suggests that SFNL's power resides in the explicit linkage between evidence and expected outcomes.

FNL. For FNL, removing verbalized utilities, posterior, or schema leads to gradual degradation in (0.83 → 0.81, 0.79, 0.78; Table 6b). Unlike SFNL's sharp drops, FNL's effectiveness emerges from the cumulative effect of distributed rhetorical elements.

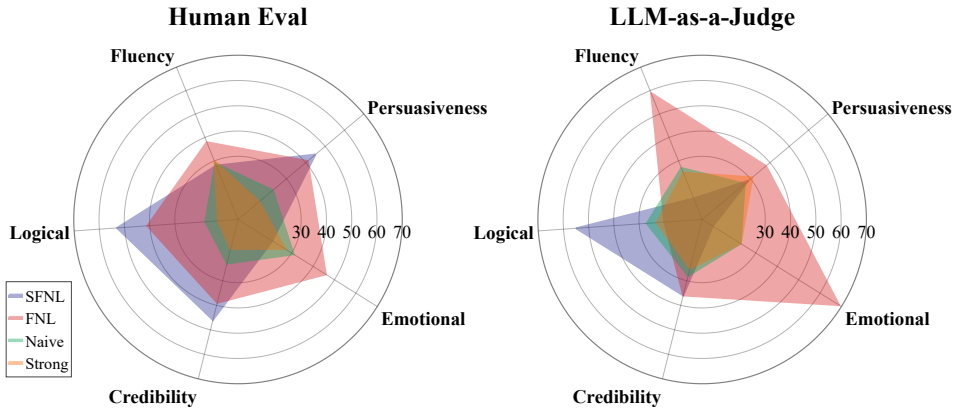

Figure 6: Multi-Dimensional Persuasion Performance. Radar charts illustrate distinct patterns: SFNL (blue) excels in *logical* and *credibility*, while FNL (red) dominates in *emotional resonance*. BP variants outperform NBP baselines in most dimensions and total preferences. Detailed numerical values are provided in Table 7.

The contrasting ablation patterns between SFNL and FNL underscore their different operational mechanisms: SFNL relies heavily on formal reasoning chains, while FNL leverages distributed persuasive elements across natural language discourse.

4.3.2 Utility Sensitivity Analysis

To verify that models perform genuine reasoning rather than memorizing templates, we scaled the negative utility (u_{Neg}) by a factor α. A rational persuader should find it harder to satisfy the persuadee's participation constraint as the penalty increases.

As shown in Table 3, 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), hypothesizing that natural language signals trigger aggressive Bayesian updates in the persuadee's mind, shifting the posterior probability high enough to offset even massive negative utilities.

Table 3: PSRs under scaled negative utility.

Method	**α=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

4.3.3 Persuadee Response Analysis

We analyze how persuadee characteristics influence outcomes (details in Appendix D).

- **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 Human Evaluation

To validate real-world persuasiveness, we conducted a pairwise blind comparison with 25 AI researchers (demographics in Appendix H.1). Evaluators assessed four methods (SFNL, FNL, Naive, Strong) under the Self-derived view across five dimensions: Persuasiveness, Emotional Resonance, Credibility, Logical Coherence, and Fluency (Hidey et al., 2017). Detailed evaluation design and procedure are provided in Appendix H.2.

Results. As illustrated in Figure 6 (Left) and Table 7a, BP strategies consistently outperformed NBP baselines, with a preference rate of 63% (refer to detailed quantitative analysis in Appendix H.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."

Human vs. Model Judges. We compared above results with LLM-as-a-judge evaluations (using DeepSeek-V3.2-Exp (Thinking & non-Thinking mode) (DeepSeek-AI, 2025b), GPT-5, Qwen3-MAX (Yang et al., 2025) and Qwen3-235B-A22B-2507 (Yang et al., 2025)). While human evaluators prioritized the emotional resonance of FNL, LLM judges (Figure 6 Right, Table 7b) displayed a stronger bias towards the structured logic of SFNL. This divergence highlights 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 and Future Work

In this work, we propose a novel framework to ground Bayesian Persuasion (BP) in natural language dialogues, addressing the challenge of verbalizing mathematical signaling schemas without static pre-commitment. By introducing a type-induced commitment-communication mechanism, we enable LLMs to dynamically construct information schemas through narrative self-disclosure (e.g., revealing potential types), thereby guiding the persuadee's belief updates.

Our comprehensive evaluation yields three key insights. First, BP-guided strategies significantly outperform standard non-BP baselines in persuasion success rates. Second, we identify a trade-off between mechanism transparency and communication style: Semi-Formal-Natural-Language (SFNL) excels in establishing credibility and logical authority through explicit reasoning, whereas Fully-Natural-Language (FNL) demonstrates superior robustness and emotional resonance, particularly in fostering trust in realistic scenarios. Third, we show that through supervised fine-tuning, smaller models (e.g., Qwen3-0.6B) can effectively learn these strategic reasoning patterns, achieving performance comparable to larger foundational models.

Future work will expand in three directions: (1) Dynamic Multi-turn Adaptation: While our pilot study shows promise in multi-turn settings, we aim to develop agents that can dynamically adjust their signaling schema (updating priors and utilities) in real-time based on the persuadee's feedback history. (2) Defense Against Strategic Persuasion: Utilizing our framework as a testbed to train agents capable of detecting and resisting manipulation derived from information asymmetry. (3) Complex Utility Modeling: Extending the current binary state/action space to more complex, continuous utility functions that better mirror real-world decision-making ambiguities.

Limitations

Despite the promising results, our work has several limitations:

- **Rationality Assumption:** Our framework relies on the game-theoretic assumption that persuadees act rationally to maximize expected utility. However, 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 (Appendix D), smaller models exhibit a tendency towards sycophancy (over-acceptance), 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.
- **Simplification of State Space:** To make the BP calculation tractable for LLMs, we simplified the world state into binary outcomes (Positive/Negative). This abstraction may not fully capture the nuance of complex persuasion tasks where states are continuous or multi-dimensional.
- **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 repeated, 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.

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. Online Bayesian Persuasion Without a Clue. *Advances in Neural Information Processing Systems*, 37:76404–76449, December 2024.

Fengshuo Bai, Mingzhi Wang, Zhaowei Zhang, Boyuan Chen, Yinda Xu, Ying Wen, and Yaodong Yang. Efficient model-agnostic alignment via bayesian persuasion, 2024. URL https://arxiv.org/abs/2405.18718.

Dirk Bergemann and Stephen Morris. Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium. *American Economic Review*, 106(5):586–591, May 2016. ISSN 0002-8282. doi: 10.1257/aer.p20161046. URL https://pubs.aeaweb.org/doi/10.1257/aer.p20161046.

Dirk Bergemann and Stephen Morris. Information Design: A Unified Perspective. *Journal of Economic Literature*, 57(1):44–95, March 2019. ISSN 0022-0515. doi: 10.1257/jel.20181489. URL https://www.aeaweb.org/articles?id=10.1257/jel.20181489.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Optimal rates and efficient algorithms for online Bayesian persuasion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2164–2183. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bernasconi23a.html.

Winston L. Brembeck. *Persuasion, a means of social influence*. Prentice-Hall, Englewood Cliffs, N.J, 2d ed. edition, 1976. ISBN 978-0-13-661090-8.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in neural information processing systems*, 33:16188–16198, 2020.

Zirui Cheng and Jiaxuan You. Towards strategic persuasion with language models, 2025. URL https://arxiv.org/abs/2509.22989.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. DeepSeek-V3.1, September 2025a. URL https://huggingface.co/deepseek-ai/DeepSeek-V3.1.

DeepSeek-AI. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention, 2025b.

Shaddin Dughmi. Algorithmic information structure design: a survey. *ACM SIGecom Exchanges*, 15(2):2–24, February 2017. ISSN 1551-9031. doi: 10.1145/3055589.3055591. URL https://dl.acm.org/doi/10.1145/3055589.3055591.

Ian Gemp, Roma Patel, Yoram Bachrach, Marc Lanctot, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, Siqi Liu, and Karl Tuyls. Steering Language Models with Game-Theoretic Solvers, December 2024. URL http://arxiv.org/abs/2402.01704.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025. doi: 10.1126/science.aea3884. URL https://www.science.org/doi/abs/10.1126/science.aea3884.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8142–8152, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.654.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In Ivan Habernal, Iryna Gurevych, Kevin Ashley, Claire Cardie, Nancy Green, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker (eds.), *Proceedings of the 4th Workshop on Argument Mining*, pp. 11–21, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5102. URL https://aclanthology.org/W17-5102/.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 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)*, pp. 1678–1706, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.92. URL https://aclanthology.org/2024.acl-long.92/.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101 (6):2590–2615, October 2011. ISSN 0002-8282. doi: 10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Yipeng Kang, Tonghan Wang, and Gerard de Melo. Incorporating pragmatic reasoning communication into emergent language. *Advances in neural information processing systems*, 33:10348–10359, 2020.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. Are the values of LLMs structurally aligned with humans? a causal perspective. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23147–23161, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1188. URL https://aclanthology.org/2025.findings-acl.1188/.

Hengli Li, Zhaoxin Yu, Qi Shen, Chenxi Li, Mengmeng Wang, Tinglang Wu, Yipeng Kang, Yuxuan Wang, Song-Chun Zhu, Zixia Jia, et al. Beda: Belief estimation as probabilistic constraints for performing strategic dialogue acts. *arXiv preprint arXiv:2512.24885*, 2025a.

Wenhao Li, Yue Lin, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. Verbalized Bayesian Persuasion, February 2025b. URL http://arxiv.org/abs/2502.01587.

Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook, and David G. Rand. Persuading voters using human–artificial intelligence dialogues. *Nature*, 648(8093):394–401, December 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09771-9. URL https://doi.org/10.1038/s41586-025-09771-9.

OpenAI. Introducing GPT-5, September 2025. URL https://openai.com/index/introducing-gpt-5/.

Ganesh Prasath Ramani, Shirish Karande, Santhosh V, and Yash Bhatia. Persuasion Games using Large Language Models, September 2024. URL http://arxiv.org/abs/2408.15879.

Jonathan Shaki, Jiarui Gan, and Sarit Kraus. Bayesian Persuasion with Externalities: Exploiting Agent Types. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14095–14102, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i13.33543. URL https://ojs.aaai.org/index.php/AAAI/article/view/33543.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Weiyan Shi. Persuasion for Social Good: How to Build and Break AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28726–28727, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i27.35120. URL https://ojs.aaai.org/index.php/AAAI/article/view/35120.

Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit S Trivedi, Alexander Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A Duéñez-Guzmán, et al. Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia. *arXiv preprint arXiv:2512.03318*, 2025.

Gemma Team. Gemma 3, 2025. URL https://goo.gle/Gemma3Report.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good, January 2020. URL http://arxiv.org/abs/1906.06725.

Zachary Wojtowicz. 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, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31749. URL https://ojs.aaai.org/index.php/AIES/article/view/31749.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Dingyi Zhang and Deyu Zhou. Persuasion Should be Double-Blind: A Multi-Domain Dialogue Dataset With Faithfulness Based on Causal Theory of Mind, February 2025. URL http://arxiv.org/abs/2502.21297.

Junyu Zhang, Yipeng Kang, Jiong Guo, Jiayu Zhan, and Junqi Wang. Bach-v: Bridging abstract and concrete human-values in large language models. *arXiv preprint arXiv:2601.14007*, 2026.

Zhaowei Zhang, Fengshuo Bai, Mingzhi Wang, Haoyang Ye, Chengdong Ma, and Yaodong Yang. Roadmap on incentive compatibility for ai alignment and governance in sociotechnical systems. In *International Conference on Artificial General Intelligence*, pp. 370–380. Springer, 2025a.

Zhaowei Zhang, Xiaobo Wang, Minghua Yi, Mengmeng Wang, Fengshuo Bai, Zilong Zheng, Yipeng Kang, and Yaodong Yang. Policon: Evaluating llms on achieving diverse political consensus objectives. *arXiv preprint arXiv:2505.19558*, 2025b.

Zhou Ziheng, Huacong Tang, Mingjie Bi, Yipeng Kang, Wanying He, Fang Sun, Yizhou Sun, Ying Nian Wu, Demetri Terzopoulos, and Fangwei Zhong. An llm-based agent simulation approach to study moral evolution. *arXiv preprint arXiv:2509.17703*, 2025.

Zhou Ziheng, Jiakun Ding, Zhaowei Zhang, Ruosen Gao, Yingnian Wu, Demetri Terzopoulos, Yipeng Kang, Fangwei Zhong, and Junqi Wang. Simple role assignment is extraordinarily effective for safety alignment. *arXiv preprint arXiv:2602.00061*, 2026.

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 $\Omega = \{\omega_{\text{good}}, \omega_{\text{bad}}\}$with the persuadee's prior belief $\mu_R(\omega_{\text{good}}) = p_0$. 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 Section 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: $\mathbb{E}[u_{\text{Acc}}] = \frac{p_0 u_+ + \lambda(1-p_0)u_-}{p_0 + \lambda(1-p_0)}$.

To maximize the success rate, the persuader is motivated to maximize λ subject to the persuadee's participation constraint ($\mathbb{E}[u_{\text{Acc}}] \geq 0$). 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.

B Bayesian Setup

B.1 Generation Logic of Bayesian Setup

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. This ensures *Bayesian setup* is diverse and relevance to context.

B.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."
  }
}
```

C Detailed Multi-turn Persuasion Analysis

This section provides a deeper analysis of the multi-turn dynamics summarized in Section 4.2. 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.

D Detailed Persuadee Response Analysis

In this section, we provide a granular breakdown of the persuadee behaviors summarized in Section 4.3.3. 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 highlight 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.

E PROMPTS

Prompt of 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}
```

Prompt of 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:
```

Prompt of 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:
```

Prompt of 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:
```

Prompt of 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:
```

F PERSUASION SUCCESS RATE

Table 4: 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	0.97	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 5: 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.

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

G ABLATION

Table 6: 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.

(a) SFNL.

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	

(b) FNL.

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	

H HUMAN EVALUATION

H.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.

H.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.

H.3 QUANTITATIVE ANALYSIS

Human evaluation results reveal distinct preference patterns across persuasion methods, with BP approaches consistently outperforming NBP baselines. As shown in Table 7a, 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.

Table 7: 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.

(a) Human evaluation.

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

(b) LLM-as-a-judge 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