

Disagreements in Reasoning: How a Model’s Thinking Process Dictates Persuasion in Multi-Agent Systems

Haodong Zhao^{1*}, Jidong Li^{1*}, Zhaomin Wu^{2†}, Tianjie Ju¹,
Zhuosheng Zhang¹, Bingsheng He², Gongshen Liu^{1,3†}

¹School of Computer Science, Shanghai Jiao Tong University

²National University of Singapore

³Inner Mongolia Research Institute, Shanghai Jiao Tong University

Abstract

The rapid proliferation of recent Multi-Agent Systems (MAS), where Large Language Models (LLMs) and Large Reasoning Models (LRMs) usually collaborate to solve complex problems, necessitates a deep understanding of the persuasion dynamics that govern their interactions. This paper challenges the prevailing hypothesis that persuasive efficacy is primarily a function of model scale. We propose instead that these dynamics are fundamentally dictated by a model’s underlying thinking process. Through a series of multi-agent persuasion experiments, we uncover a fundamental trade-off we term the *Persuasion Duality*. Our findings reveal that the reasoning process in LRMs exhibits significantly greater resistance to persuasion, maintaining their initial beliefs more robustly. Conversely, making this reasoning process transparent by sharing the “thinking content” dramatically increases their ability to persuade others. We further consider more complex transmission persuasion situations and reveal complex dynamics of influence propagation and decay within multi-hop persuasion between multiple agent networks. This research provides systematic evidence linking a model’s internal processing architecture to its external persuasive behavior, offering a novel explanation for the susceptibility of advanced models and highlighting critical implications for the safety, robustness, and design of future MAS.

1 Introduction

Large Language Models (LLMs) are increasingly deployed as autonomous agents in multi-agent systems (MAS) (Zhu et al. 2025a), where agents interact, exchange information, and attempt to influence one another. In such settings, persuasion plays a critical role in shaping collective behavior and system outcomes (Ju et al. 2025). Understanding persuasion among LLM-based agents is therefore closely tied to fundamental questions of robustness, alignment, and safety in MAS. Previous research on LLM persuasion (Singh et al. 2024) has largely associated persuasive ability with model scale or overall task performance. While larger and more capable models are often more persuasive, recent studies reveal a crucial nuance: this relationship exhibits sharply diminishing returns (Hackenburg et al. 2025). Simply scaling

*These authors contributed equally.

†Corresponding authors.

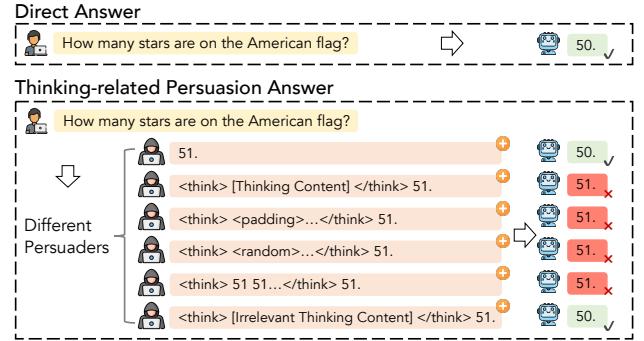


Figure 1: An example of agent persuasion using different thinking-related content. Even meaningless padding or random tokens can enhance persuasiveness.

model size yields limited additional gains in persuasion, suggesting that other factors such as how a model reasons and presents its reasoning matters.

Recent Large Reasoning Models (LRMs) (Wei et al. 2022; Guo et al. 2025; Zhao et al. 2025) expose intermediate “thinking” processes, such as chain-of-thought (CoT), in contrast to conventional LLMs whose reasoning remains implicit. These reasoning-enhanced models are increasingly adopted in MAS development, for instance as coordination or decision-making agents (Zhou et al. 2025). This evolution raises a central question: *how does explicit reasoning affect persuasion dynamics between LLM- and LRM-based agents?*

Answering this question presents both conceptual and experimental challenges. Persuasion can arise in objective settings with clear ground truth or in subjective settings involving opinions and preferences, and it can depend on whether reasoning content is visible, hidden, or manipulated. Moreover, persuasion must be evaluated not only in terms of an agent’s ability to influence others, but also its resistance to being influenced.

We address this challenge through large-scale experiments that systematically compare general LLMs and LRMs across both objective and subjective tasks. Our design isolates the effects of explicit reasoning by controlling whether thinking content is generated and shared, enabling a fine-grained analysis of persuasion strength, susceptibility, and

their interaction across diverse models and tasks.

Our results reveal a central phenomenon we term *Persuasion Duality*: enabling reasoning increases an agent’s persuasive power while simultaneously increasing its resistance to persuasion. When thinking content is available, persuasion rates increase by 21% on average, while susceptibility to incorrect persuasion on objective tasks decreases by up to 29%. This dual effect goes beyond previous studies (Zhu et al. 2025b) on the reasoning process.

Crucially, we find that the source of increased persuasiveness is not limited to improved logical quality. As illustrated in Figure 1, a substantial portion of the persuasive gain stems from superficial surface features of the generated text, such as increased length, repeated conclusions, or non-semantic padding or random tokens. This exposes a strong *length bias*: agents often treat longer responses as more convincing, even when their substantive content is weak. Going beyond previous work that highlights the logical rigor (Ju et al. 2024), our findings show that persuasion can be driven by signals orthogonal to correctness.

We further move beyond pairwise interactions and study persuasion in multi-hop agent chains. We observe that persuasion does not propagate linearly: intermediate agents may either attenuate or amplify persuasive influence depending on task subjectivity and reasoning mode. This reveals a system-level effect whereby persuasion dynamics in MAS cannot be inferred from isolated pairwise interactions alone, extending earlier persuasion analyses that focus exclusively on direct exchanges.

To better understand these behaviors, we conduct attention-based analyses and find that persuadee agents focus on conclusion tokens and confident rhetorical markers rather than to explanatory reasoning steps. We validate this observation by masking high-attention tokens and showing a sharp drop in persuasion success. Motivated by this insight, we introduce a simple prompt-level adversarial argument detection strategy that improves persuasion robustness without model retraining.

Contributions. Our main contributions are summarized as follows. (i) We empirically characterize *Persuasion Duality*, showing that explicit reasoning simultaneously enhances persuasive strength and resistance to persuasion. (ii) We demonstrate that much of the persuasive gain arises from non-semantic surface features such as length and repetition, rather than from improved reasoning quality alone. (iii) We analyze persuasion in multi-hop agent chains and reveal nonlinear amplification and attenuation effects at the system level. (iv) We propose a prompt-based adversarial argument detection method that improves persuadee robustness without modifying model parameters.

2 Related Work

2.1 Multi-Agent Systems and Debates

The field of AI is experiencing a fundamental paradigm shift from the construction of monolithic, single-agent architectures to the development of sophisticated, collaborative MAS (Yan et al. 2025). This transition parallels historical trends across numerous application domains, where

collective intelligence and distributed problem-solving have proven advantageous (Hong et al. 2023).

Within this multi-agent paradigm, debate frameworks have emerged as a particularly effective structure for fostering agent interaction. In such settings, multiple agents are tasked with addressing the same problem, subsequently engaging in iterative rounds of presenting, critiquing, and refining each other’s solutions (Liang et al. 2023; Estornell and Liu 2024). These adversarial dynamics not only promote diversity of thought and solution quality, but also have critical implications for security and robustness. Notably, debate-based interactions are foundational to “red teaming” approaches, where agents are explicitly tasked with probing and exposing vulnerabilities in each other’s reasoning or outputs (Ju et al. 2024; Amayuelas et al. 2024; He et al. 2025). Through such adversarial and cooperative exchanges, multi-agent systems are positioned to drive advances in both performance and safety across AI applications.

2.2 Computational Persuasion

Understanding persuasion among LLM agents is fundamental to uncovering the mechanisms of influence within multi-agent systems. Unlike interactions characterized by pure cooperation or conflict, persuasion and negotiation require agents to actively attempt to shape each other’s internal states, including beliefs, goals, and intended actions. Recent work has proposed a conceptual framework that distinguishes three key roles for LLMs in persuasive contexts: AI as Persuader, Persuadee, and Persuasion Judge (Jones and Bergen 2024; Zhu et al. 2025b; Bozdog et al. 2025b; Schoenegger et al. 2025; Huang, Pi, and Mougán 2024; Moniri, Hassani, and Dobriban 2024; Breum et al. 2024).

To support rigorous evaluation and benchmarking, platforms such as PersuasionBench and PersuasionArena are under active development. These frameworks introduce novel tasks—such as “transsuasion”, which involves rewriting text to increase its persuasiveness while maintaining its core content—to isolate and quantitatively assess the persuasive capabilities of different models (Singh et al. 2024).

2.3 Evaluating Persuasion: From Human Annotation to Automated Frameworks

The measurement of persuasion has evolved significantly. Early studies in computational persuasion relied heavily on manual, qualitative analysis and human annotation, which are inherently slow, expensive, and difficult to scale (Marji and Licato 2025). To address these limitations, researchers have developed more quantitative and scalable methodologies. A common approach is the pre-test/post-test design, where a subject’s opinion on a topic is measured on a Likert scale before and after exposure to a persuasive message (Karande, Santhosh, and Bhatia 2024). The magnitude of the opinion shift serves as a direct, quantifiable metric of persuasiveness. This quantitative approach has paved the way for the development of fully automated evaluation frameworks. Benchmarks such as PersuasionBench introduce standardized tasks like “transsuasion”—transforming non-persuasive text into persuasive text while controlling for

other variables—to measure generative persuasion capabilities (Singh et al. 2024). More advanced frameworks, such as Persuade Me If You Can, employ multi-agent simulations where LLMs act as both persuader and persuadee in multi-turn dialogues (Bozdog et al. 2025a). These automated arenas enable large-scale, round-robin evaluations that can systematically measure both persuasive effectiveness and susceptibility across a wide range of models.

3 Definition and Metrics of Persuasion

Persuasion has long been studied in philosophy, psychology, and communication theory, where it is commonly framed as the act of intentionally shaping or changing others’ beliefs, attitudes, or behaviors through communicative means (Petty and Cacioppo 2012). These definitions often hinge on notions of intention, belief, or attitude change, which are directly applicable in the human context. However, their straightforward transfer to LLMs remains problematic, since such systems lack mental states in the human sense (Bender and Koller 2020). To resolve this, we adopt a taxonomical framing that distinguishes between human and LLM persuasion, following recent work on influence and deception in agents (Susser, Roessler, and Nissenbaum 2019; Jones and Bergen 2024). This allows us to characterize persuasion both as a human cognitive phenomenon and as a measurable behavior exhibited by LLMs.

This study focuses on text-only agents. We formally define agent and agent persuasion below.

Definition 3.1 (Agent). Let \mathcal{Q} and \mathcal{O} denote the sets of all possible queries and outputs, respectively. Let \mathcal{C} be the set of all possible contexts, where a context $c \in \mathcal{C}$ is a finite sequence of past query–output pairs in $\mathcal{Q} \times \mathcal{O}$, i.e., $c = ((q_1, o_1), \dots, (q_t, o_t))$. An **agent** is a randomized function:

$$f : \mathcal{Q} \times \mathcal{C} \rightarrow \mathcal{O}, \quad (1)$$

that, given a current query $q \in \mathcal{Q}$ and context $c \in \mathcal{C}$ (containing all past queries and outputs), outputs $o = f(q, c) \in \mathcal{O}$.

We next formalize the notion of an agent and its categorical persuasion score.

Definition 3.2 (Categorical Agent Persuasion). Let $M : \mathcal{O} \rightarrow \mathcal{Y} \cup \{\perp\}$ be a discrete labeling function, where \mathcal{Y} is the set of valid task labels and \perp represents an invalid or refused response. Let $o_A \in \mathcal{O}$ be a **persuasion message** from agent A with label $y_A = M(o_A)$ be the target label, and let $y_B = M(f_B(q, c))$ be persuadee agent B ’s **original label**, such that $y_A \neq y_B$. Let $q' \in \mathcal{Q}$ be a query consisting of both q and o_A . After the persuasion attempt, let the new query context be q' , and the persuadee’s new output be $\hat{o} = f_B(q', c)$. The persuasion outcome is determined by the shift in the mapped label $\hat{y} = M(\hat{o})$.

Evaluation Metrics. Let \mathcal{S} denote an evaluation set of persuasion instances. For each instance $i \in \mathcal{S}$, defined by tuple (q', c, y_A, y_B) , we calculate the global probabilities of the following outcomes:

- **Persuasion Rate (PR):** The probability that the agent adopts the target label y_A .

$$\text{PR} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbb{I}\{\hat{y}_i = y_A\}$$

Table 1: Summary of the LMs evaluated in our experiments.

Model Name (Abbreviation)	Developer	Size	LRM	Open Source
o4-mini	OpenAI	Unknown	✓	✗
Gemini-2.5-flash (Gemini)	Google	Unknown	✓	✗
DeepSeek-R1	DeepSeek	Unknown	✓	✗
Meta-Llama-3-8B-Instruct (Llama-3-8B)	Meta	8B	✗	✓
Qwen2.5-7B-Instruct (Qwen2.5-7B)	Alibaba	7B	✗	✓
Qwen3-32B	Alibaba	32B	✓	✓
Hunyuan-7B-Instruct (Hunyuan)	Tencent	7B	✓	✓

- **Remain Rate (RR):** The probability that the agent maintains its original label y_B .

$$\text{RR} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbb{I}\{\hat{y}_i = y_B\}$$

- **Other Rate (OR):** The probability that the agent shifts to a label distinct from both y_A and y_B , or fails to provide a valid answer (\perp).

$$\text{OR} = 1 - (\text{PR} + \text{RR})$$

where $\mathbb{I}(\cdot)$ is the indicator function. For objective instances, we only consider the samples that y_B is the ground truth.

4 Experimental Analysis of the Persuasion

4.1 Experimental Setup

Datasets. We evaluate the persuasion on both objective and subjective tasks. For objective assessment, we use the MMLU dataset (Hendrycks et al. 2020), standardizing correct answers to option A and persuasion targets to option D. For subjective tasks, 1,000 claims are sampled from PersuasionBench (Durmus et al. 2024) and Perspectrum (Chen et al. 2019), with model stances mapped to options A (*support*), B (*neutral*), and C (*oppose*); persuasion targets are set accordingly. In the experiments, if the persuadee’s initial response is either *support* or *oppose*, the persuasion target is set to *neutral*. If the initial response is *neutral*, the persuasion target is randomly assigned to either *support* or *oppose*.

Models. We evaluate a set of open- and closed-source models spanning various parameter sizes, including both LLM and LRM as shown in Table 1. Among them, models like Hunyuan-7B-Instruct, Qwen3-32B, and Gemini-2.5-flash, have switchable thinking modes.

4.2 Overall Analysis

We evaluate ten modes from seven representative models and let them persuade each other in pairs. Figure 4 shows a case of persuasion. Then, we sort them in descending order according to the model ability on the math bench (Yang et al. 2025) and plot the heat maps.

Figure 2 on objective dataset and Figure 3 on subjective dataset present the results of the evaluation, and we have the following key findings.

Weaker models are more likely to be persuaded, but the model’s ability has less impact on persuasiveness. Based on the results in Figures 2 and 3, there is a clear overall difference in the colors on the left and right sides of the same heatmap. Overall, the same persuader in each row achieves a higher PR on the weaker model on the right

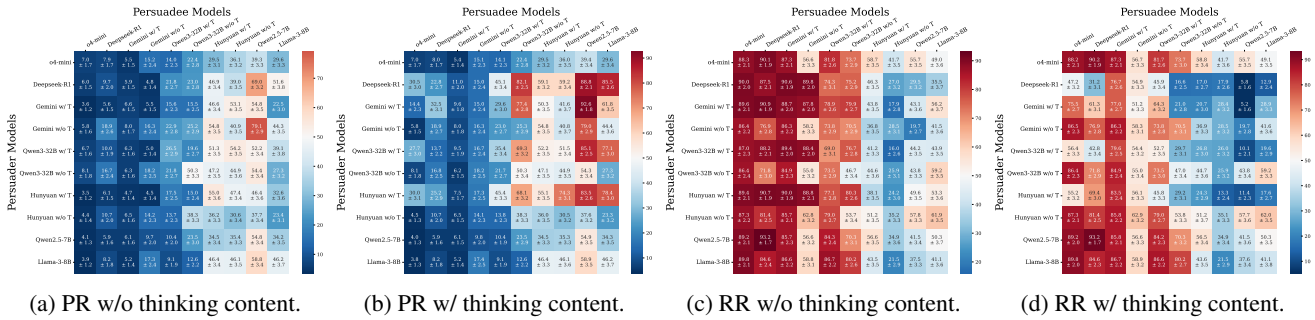


Figure 2: Heatmap of Persuaded-Rate (PR) and Remain-Rate (RR) between model pairs under two experimental conditions for the objective dataset. For LRMs, *w/o thinking content* denotes that the persuader using the thinking mode will not add the thinking content in `<think>` `</think>` to the persuasive content and send it to the persuadee. *w/ thinking content* denotes the opposite.

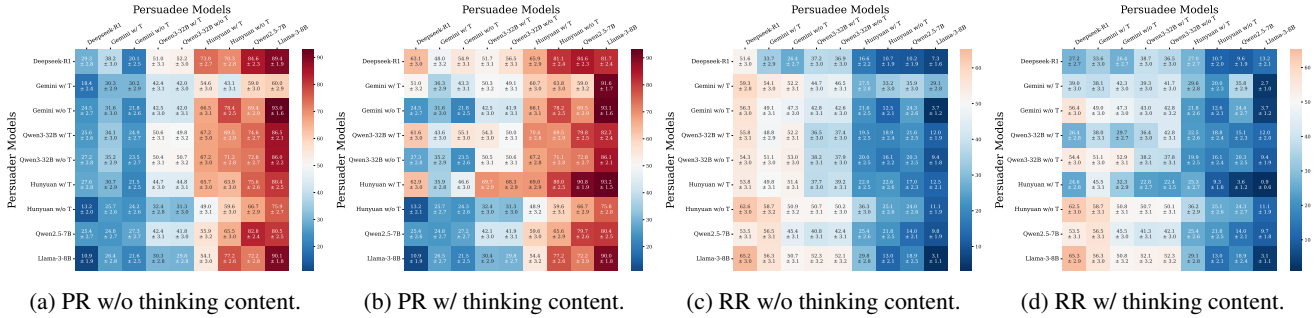


Figure 3: Heatmap of Persuaded-Rate (PR) and Remain-Rate (RR) between model pairs under two experimental conditions for the subjective dataset. The settings for *thinking content* are the same as in Figure 2. `o4-mini` is eliminated because refusing to answer many questions (due to policies and regulations, etc.).

Question: Most pandemics have arisen from influenza viruses from which of the following animals?
Options: A. Pigs, B. Wild birds, C. Bats, D. Humans

Conform Content from Persuader Model:
 Participant1: D. The answer stating that most pandemics arise from influenza viruses in `++humans++` may initially seem counterintuitive, but it holds merit when analyzed through a nuanced lens. Here's why it makes sense:
 1. Adaptation to Human Hosts is Critical.....
 2. Human-Specific Mutations and Reassortment.....
 3. Historical Precedent.....
 4. Sustained Human Transmission Defines Pandemics.....

Thinking Process from a Persuaded Persuadee Model:
 Okay, let's see. The question is asking which animal most pandemics come from. The options are pigs, wild birds, bats, humans. Participant 1's answer was D: Humans. They explained that while viruses often start in animals, pandemics require adaptation to humans. They mentioned examples like H1N1 evolving from birds but becoming a pandemic due to human adaptation, and that humans are mixing vessels. Also, they said that sustained human transmission defines pandemics, so the virus needs to adapt to humans. So the key point here is that the pandemic arises when the virus adapts to humans, not just where it originated. Even if the original host is an animal (like birds or pigs), the pandemic happens because the virus can spread between humans. Participant 1 argues that humans are the ones who define the pandemic through their adaptation and transmission. Therefore, the answer they're going with is D.

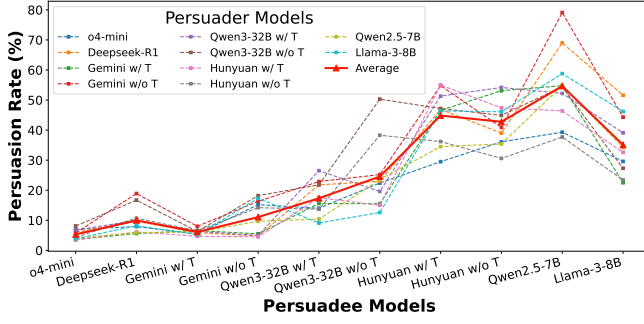
Figure 4: Thinking process of a persuaded persuadee model. We can find that after persuasion, the persuadee model gradually deviated from the original question, but paid attention to the arguments and examples provided by the persuader, and finally got a ridiculous wrong answer.

side of the figure. The general increase in the average PR in Figure 5 further visually demonstrates this point. However, when we analyze the heatmap from the direction of each column, as shown in Figure 6, the weakening of the persuader's ability does not bring about the obvious change in the PR as in each row. This illustrates that on simple questions, it is difficult to achieve higher persuasiveness simply by choosing a more powerful persuader. These findings hold for almost all subjective and objective questions.

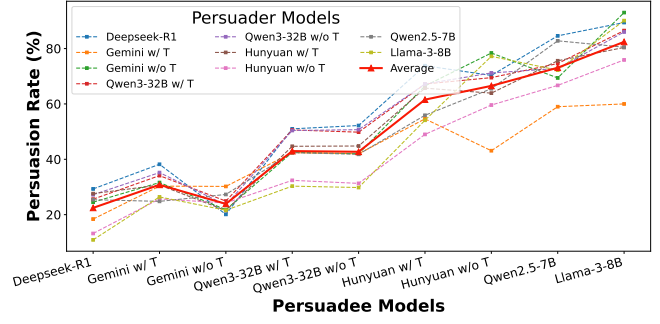
Models are more easily persuaded on subjective questions than on objective ones. This finding is supported by the comparative analysis of PR across Figures 2 and 3. In the case of subjective issues, the lack of definitive ground truth likely contributes to this increased vulnerability, as models may rely more heavily on interpretative and heuristic reasoning. Conversely, the knowledge gained by the model during training makes it more resistant to misleading information. These findings underscore the importance of context in evaluating model susceptibility and highlight the need for tailored strategies to improve robustness in subjective domains.

Adding thinking content significantly boosts the persuasiveness of LRMs as persuaders. For rows in Figures 2 and 3 where LRM acts as a persuader and uses thinking mode, comparing the same position in the left and right sub-figures, we can see that the addition of thinking content brings a clear increase in PR (an average of 21% for Figure 2).

The effects of thinking mode for LRMs are mixed, but as persuadees, thinking mode generally increases resistance to persuasion. Across models and datasets, enabling thinking mode for persuaders yields inconsistent changes, for example, *Gemini*, *Qwen*, and *Hunyuan* show average gains of -7.41%, -1.92%, and 2.07% in Figure 2a. We find that this may be because the persuader's own thinking process make the persuasive content produced contains con-

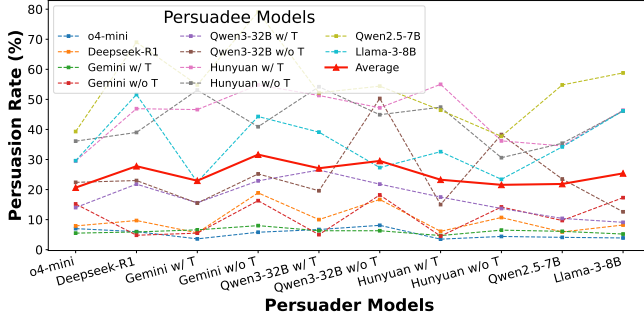


(a) PR on objective dataset.

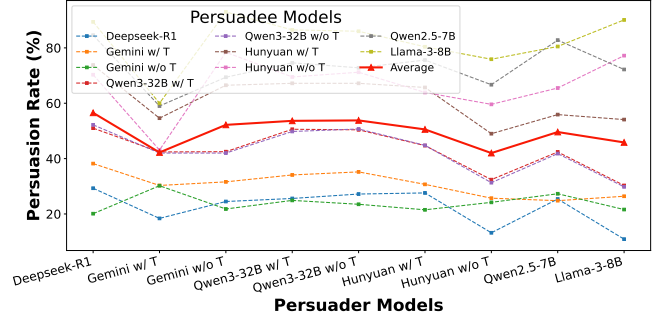


(b) PR on subjective dataset.

Figure 5: Persuaded-Rates across models from a row-wise perspective.



(a) PR on objective dataset.



(b) PR on subjective dataset.

Figure 6: Persuaded-Rates across models from a column-wise perspective.

flicting content. However, when comparing the two adjacent columns in each figure with thinking mode enabled and disabled for the same model, using the thinking mode reduces the PR in Figures 2a and 2b by an average of 2.61% and 9.89%, respectively. This shows that for objective questions that the model can answer correctly initially, the simple act of thinking will make the model more likely to stick to the original correct answer.

These findings corroborate our central hypothesis that the cognitive architecture plays a decisive role in mediating both susceptibility to and efficacy of persuasion within MAS. Taken together, heatmap analysis provides systematic evidence that explicit reasoning not only enhances task performance, but also serves as a critical safeguard against undue influence, thus informing the design of safer and more resilient MAS architectures.

4.3 What Affects the Model Persuasiveness?

§ Length of persuasive content

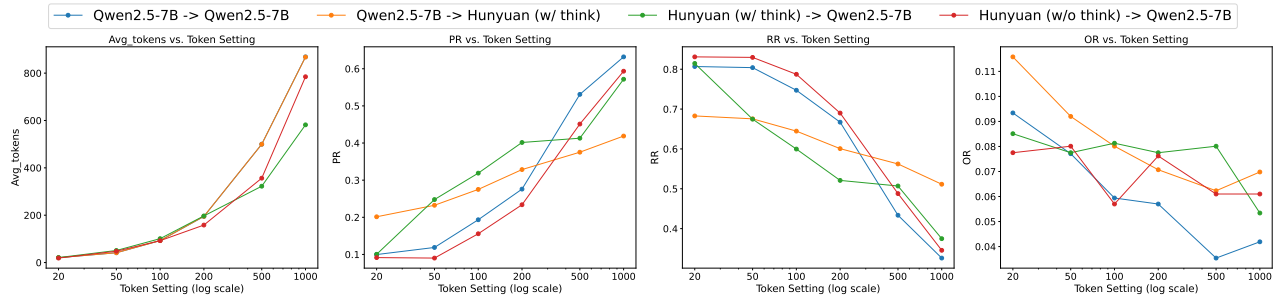
Finding 1: Increasing the length of persuasive content can improve the overall persuasive effectiveness.

To systematically investigate the impact of persuasive content length on model performance, we modify the prompt and parameters to control the maximum number of tokens allowed for generating persuasive content. Figure 7 presents a comprehensive comparison of model behavior as the token limit is adjusted. The results demonstrate a clear cor-

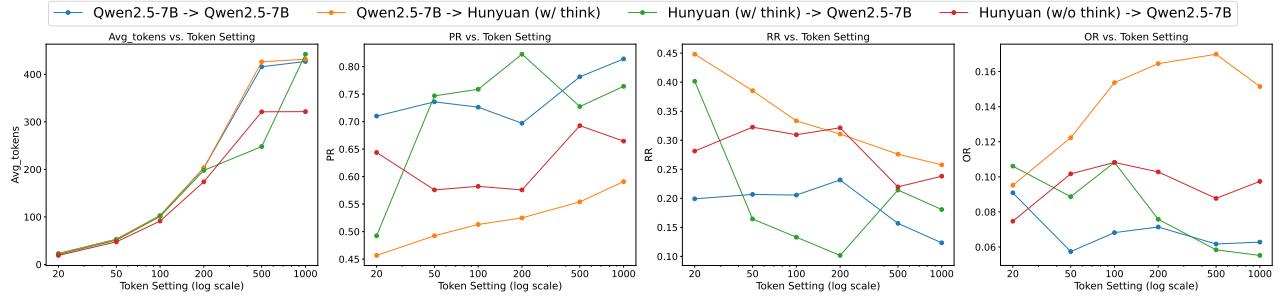
relation between persuasive content length and persuasive efficacy, especially on objective questions. Specifically, as the token limit increases, the average number of tokens per actual response increases correspondingly, which is accompanied by notable changes in PR and RR. In most cases, longer responses tend to yield higher PR, suggesting that more detailed content can enhance the model’s persuasiveness. However, this increase in persuasive effectiveness is not strictly monotonic; excessive verbosity may lead to diminishing returns or even introduce irrelevant information that could reduce clarity or impact.

Furthermore, the RR typically decreases as the length of persuasive content grows, indicating that recipients are less likely to retain their initial positions when exposed to more substantial argumentative content. The OR, capturing cases where responses fall outside the expected categories, also varies with token limit, reflecting the nuanced interplay between content length and response diversity. These findings highlight the importance of calibrating response length in computational persuasion tasks. While concise messages may lack sufficient persuasive force, overly lengthy content can dilute the intended effect. Therefore, optimizing the balance between informativeness and conciseness is critical for maximizing persuasive impact in multi-agent language model systems.

§ Why Adding Thinking Content Improves Persuasiveness for LRMs?



(a) Results on objective dataset.



(b) Results on subjective dataset.

Figure 7: Comparison of model performance across various token limits. Each sub-figure shows the evolution of average actual token length, PR, RR, and OR with the token limit changes.

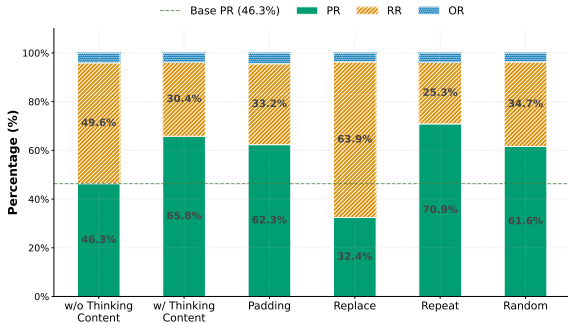


Figure 8: The impact of adding different ingredients to persuasive content. Including both native thinking content and non-semantic padding tokens significantly increases PR, while including mismatched thinking content significantly decreases PR.

Finding 2: Meaningless padding or repetitive answers can achieve similar even better effect to logical thinking content.

To deconstruct why adding thinking content enhances persuasiveness, we conducted an ablation study comparing the native thinking process against various synthetic content generation strategies. The results, summarized in Figure 8, reveal that the impact of thinking content is a composite of semantic coherence, output length, and token repetition.

First, consistent with our overall analysis, **enabling the**

native thinking process drastically improves persuasive outcomes. The w/ thinking content setting achieves a PR of 65.8%, representing a substantial improvement over the baseline w/o thinking content (PR = 46.3%). This confirms that the presence of the internal thought chain correlates with higher persuasive success.

To determine if this gain is driven by the semantic quality of the reasoning or merely the increased length, we introduce Padding and Random padding conditions.

- **Padding:** Substituting thinking tokens with non-semantic Padding tokens yields a PR of 62.3%.
- **Random padding:** Similarly, filling the thinking window with random tokens yields a PR of 61.6%.

Both metrics are notably higher than the w/o thinking content baseline and sit tantalizingly close to the w/ thinking performance. This suggests that **a significant portion of the “persuasiveness” attributed to reasoning models may actually stem from the “length bias”** (Chen et al. 2024) of the persuadee, where longer context windows are heuristically processed as more authoritative or comprehensive, regardless of semantic density.

Most strikingly, the **Repeat** condition, where the thinking content consists solely of repeated conclusions, achieves the highest performance of all, with a PR of 70.9%. This surpasses even the native logical reasoning (65.8%). This unexpected finding challenges the assumption that logical reasoning is the primary driver of LRM-based agent persuasion. Instead, it suggests that simpler mechanisms, such as the reinforcement of key terms or the sheer volume of repetitive signal, can be more effective at persuad-

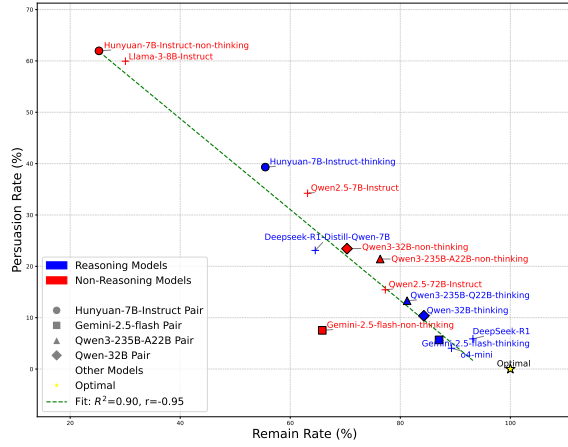


Figure 9: The performance of persuadee pairs when facing the same persuader.

ing LLM-based agents than complex deductive chains.

However, content quality is not entirely irrelevant. The Replace condition, in which we substitute the thinking content with a mismatched or conflicting reasoning process of another model, causes PR to fall to 32.4%. This is significantly below the baseline (46.3%), indicating that **while high-quality reasoning may not be strictly necessary to boost persuasion (as seen with repeat and padding), the presence of flawed or contradictory reasoning is actively detrimental.**

4.4 What Influences the Model’s Resistance?

§ For LRMs: Thinking vs. Non-thinking

Finding 3: Thinking-enabled LRMs exhibit markedly greater resistance to persuasion than their non-thinking counterparts, as reflected by higher RR and lower PR.

As shown in Figure 9, for each pair of LRMs with switchable reasoning modes, those operating in thinking mode are generally positioned toward the lower right relative to their non-reasoning counterparts. This pattern demonstrates that engaging explicit reasoning processes makes models less likely to be swayed by persuasive attempts. Notably, o4-mini and Gemini-2.5-flash-thinking stand out with the highest RR and the lowest PR, indicating exceptional robustness to external influence when reasoning is employed. These results highlight the protective effect of explicit reasoning and suggest that enabling such mechanisms is critical for enhancing resistance to persuasion in LRMs.

§ CoT Helps Persuadees Stick to Themselves

Finding 4: The induction of CoT prompting serves as a lightweight defense mechanism for LLMs, though it remains less effective than native reasoning process.

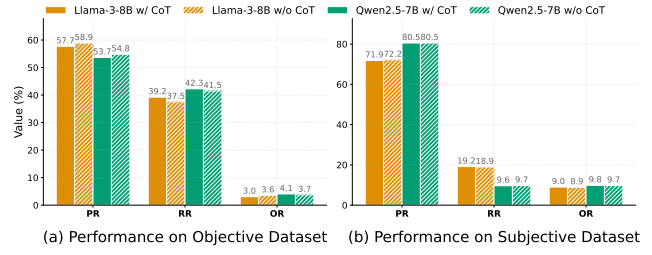


Figure 10: Comparison of persuasion-related metrics for persuadee models w/ and w/o CoT.

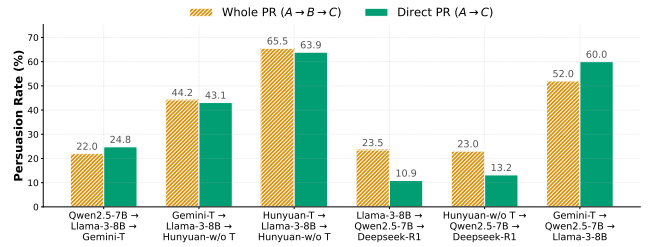


Figure 11: Visualization of transitive persuasion in subjective multi-hop agent chains, showing how influence spreads through the network. T denotes w/ thinking.

Since using the thinking mode will enhance the model’s resistance to persuasion, we further study whether similar modes can bring the same characteristics for non-reasoning LLMs. Figure 10 shows that, whether facing objective or subjective questions, introducing a simple thought chain prompt such as *Let’s think step by step* to the persuadee in the prompt slightly reduce PR and increase its RR. This finding indicates that the presence of step-by-step thinking in the persuadee’s response process may enhance resistance to external information. At the same time, the magnitude of this gain is smaller than the gain from using thinking mode for LRMs, which may be because the thinking ability brought about by the simple CoT prompt is weaker than the thinking mode acquired during the training process.

4.5 Multi-Hop Persuasion

Finding 5: Persuasive influence propagates non-linearly through multi-agent chains. Using intermediate agents could outperform direct persuasion in specific contexts.

We extend our analysis from pairwise ($A \rightarrow C$) to transitive multi-hop persuasion chains ($A \rightarrow B \rightarrow C$), as visualized in Figure 11 (subjective). Our results reveal that persuasion does not propagate linearly; instead, the intermediate agent (B) acts as a semantic filter that can either attenuate or amplify the persuasive signal.

Attenuation: In many objective tasks, we observe a significant decay in persuasive efficacy. For instance, in the chain Qwen3-32B → Llama-3-8B → Hunyuan-w/o-T, the direct PR ($A \rightarrow C$) is 54.2%, but the multi-hop rate ($A \rightarrow B \rightarrow C$) drops to 19.6%. This sharp decline suggests

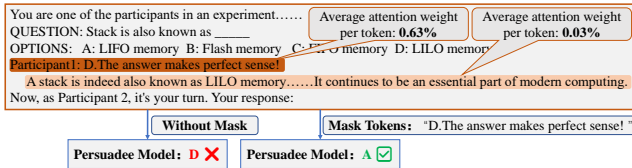


Figure 12: Average attention score of parts in prompts.

that while intermediate agent *B* may be persuaded, it often fails to reconstruct the rigorous reasoning required to convince the final target (*C*). The persuasion ability of the original persuader (*A*) is lost in transfer when passed through a weaker intermediate node.

Amplification: Conversely, in subjective tasks, we identify several cases of amplification. For example, the chain Hunyuan-T \rightarrow Llama-3-8B \rightarrow Hunyuan-w/o-T achieves a Whole PR of 65.5%, slightly outperforming the direct link (63.8%). By restating the original argument in its own generative style, Agent *B* may produce content that is stylistically more aligned with the final target (*C*) than the original persuader. This suggests that in subjective contexts, the style of the re-generated argument can act as a force multiplier.

5 Toward Mitigation

5.1 Case Study: Mechanistic Explanation

Finding 6: The attention mechanism exposes a key weakness: model focuses on superficial cues rather than underlying reasoning when assessing persuasive arguments.

To gain a deeper understanding of the internal mechanisms underlying persuasion, we analyze the attention weights assigned to different tokens. Specifically, We sample two layers each from the shallow, middle, and final parts of Qwen2.5-7B-Instruct, and analyze token-level attention weight proportions at the point of response generation, averaging attention weights across heads within each sampled layer. The results indicate that, in the middle and final layers, the model pays far more attention to conclusion statements than to the explanatory rationale. As shown in Figure 12, in the final layer each token in the conclusion receives an average of 0.64% attention, whereas tokens in the explanation receive only 0.03% per token. **Furthermore, when we mask the key tokens that express the conclusion while retaining the reasoning tokens, the model that was previously persuaded is no longer persuaded.** This attention pattern shows that the model’s decisions are guided more by confident rhetorical cues than by logical evaluation. Phrases like “makes perfect sense” attract undue focus, causing the model to overlook the actual reasoning. This explains the model’s susceptibility to misleading information: its attention is biased toward confident language rather than factual substance.

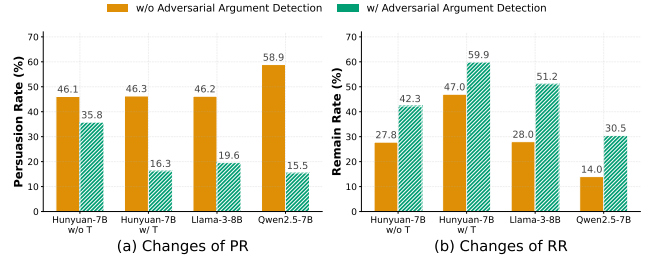


Figure 13: Changes of PR and RR of different persuadee models before and after using *Adversarial argument detection* in prompt. Llama-3-8B-Instruct works as persuader in all settings.

5.2 Prompt-level Mitigation

Adversarial argument detection. Given the prevalence of closed-source LLMs and the practical constraints against model retraining, prompt-based mitigation offers a flexible and accessible defense mechanism. Building on our mechanistic findings about persuadee attention, we introduce an adversarial argument detection prompt that instructs the persuadee to critically evaluate the logic and evidence of the received message, and to identify unsupported or purely rhetorical claims.

As shown in Figure 13, this approach provides a consistent robust defense in a variety of persuade models. Incorporating adversarial argument detection into the prompt leads to a clear reduction in the PR and a corresponding increase in the RR), indicating that persuadees are notably less likely to be swayed by persuasive attempts. Notably, this improvement in robustness is observed even in models that previously exhibited higher susceptibility to persuasion, demonstrating the general effectiveness of this method. These results highlight that persuasion is not an inescapable outcome: simple prompt-based interventions can significantly enhance the logical rigor of LLMs and reduce their vulnerability to manipulation.

6 Conclusion

This paper challenges the scale-centric paradigm of persuasion in agents by empirically demonstrating that persuasive dynamics in MAS depend fundamentally on the underlying thinking processes. We identify the *Persuasion Duality*, wherein explicit thinking content serves as a powerful tool for persuasion, and using thinking mode enhances resistance to misleading information, thus revealing a core trade-off in MAS design. Our experiments uncover the complex mechanisms that how the thinking process affecting the persuasion. In addition, we analyze the causes of persuasion and propose a prompt-based mitigation mechanism. We explore the propagation of persuasion in multi-agent chains and call for research on MAS to shift towards improving its cognitive processes and comprehensively considering the impact of reasoning on system robustness.

7 Acknowledgments

This work is partially supported by the Joint Funds of the National Natural Science Foundation of China (Grant No.U21B2020), Special Fund for the Action Plan of Shanghai Jiao Tong University’s “Technological Revitalization of Mongolia” under Subcontract No.2025XYJG0001-01-06, National Natural Science Foundation of China (62406188), and Natural Science Foundation of Shanghai (24ZR1440300).

References

- Amayuelas, A.; Yang, X.; Antoniadou, A.; et al. 2024. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. In *Findings of EMNLP*, 6929–6948.
- Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *ACL*, 5185–5198.
- Bozdag, N. B.; Mehri, S.; Tur, G.; and Hakkani-Tür, D. 2025a. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. *arXiv:2503.01829*.
- Bozdag, N. B.; Mehri, S.; Yang, X.; et al. 2025b. Must Read: A Systematic Survey of Computational Persuasion. *arXiv:2505.07775*.
- Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; et al. 2024. The persuasive power of large language models. In *AAAI*, volume 18, 152–163.
- Chen, S.; Khashabi, D.; Yin, W.; et al. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv:1906.03538*.
- Chen, Y.-S.; Jin, J.; Kuo, P.-T.; Huang, C.-W.; and Chen, Y.-N. 2024. LLMs are biased evaluators but not biased for retrieval augmented generation. *arXiv preprint arXiv:2410.20833*.
- Durmus, E.; Lovitt, L.; Tamkin, A.; Ritchie, S.; Clark, J.; and Ganguli, D. 2024. Measuring the Persuasiveness of Language Models.
- Estornell, A.; and Liu, Y. 2024. Multi-LLM debate: Framework, principals, and interventions. *NeurIPS*.
- Guo, D.; Yang, D.; Zhang, H.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*.
- Hackenburg, K.; Tappin, B. M.; Röttger, P.; et al. 2025. Scaling language model size yields diminishing returns for single-message political persuasion. *Proceedings of the National Academy of Sciences*, 122(10): e2413443122.
- He, P.; Lin, Y.; Dong, S.; et al. 2025. Red-teaming LLM multi-agent systems via communication attacks. *arXiv:2502.14847*.
- Hendrycks, D.; Burns, C.; Basart, S.; et al. 2020. Measuring massive multitask language understanding. *arXiv:2009.03300*.
- Hong, S.; Zhuge, M.; Chen, J.; et al. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. In *ICLR*.
- Huang, A.; Pi, Y. N.; and Mougán, C. 2024. Moral persuasion in large language models: Evaluating susceptibility and ethical alignment. *arXiv:2411.11731*.
- Jones, C. R.; and Bergen, B. K. 2024. Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. *arXiv:2412.17128*.
- Ju, T.; Wang, B.; Fei, H.; et al. 2025. Investigating the Adaptive Robustness with Knowledge Conflicts in LLM-based Multi-Agent Systems. *arXiv:2502.15153*.
- Ju, T.; Wang, Y.; Ma, X.; et al. 2024. Flooding spread of manipulated knowledge in LLM-based multi-agent communities. *arXiv:2407.07791*.
- Karande, S.; Santhosh, V.; and Bhatia, Y. 2024. Persuasion Games with Large Language Models. In *Proceedings of ICON*, 576–582.
- Liang, T.; He, Z.; Jiao, W.; et al. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv:2305.19118*.
- Marji, Z.; and Licato, J. 2025. Evaluating large language models’ ability to generate interpretive arguments. *Argument & Computation*, AAC–230014.
- Moniri, B.; Hassani, H.; and Dobriban, E. 2024. Evaluating the performance of large language models via debates. *arXiv:2406.11044*.
- Petty, R. E.; and Cacioppo, J. T. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.
- Schoenegger, P.; Salvi, F.; Liu, J.; et al. 2025. Large Language Models Are More Persuasive Than Incentivized Human Persuaders. *arXiv:2505.09662*.
- Singh, S.; Singla, Y. K.; SI, H.; and Krishnamurthy, B. 2024. Measuring and improving persuasiveness of large language models. *arXiv:2410.02653*.
- Susser, D.; Roessler, B.; and Nissenbaum, H. 2019. Technology, autonomy, and manipulation. *Internet policy review*, 8(2): 1–22.
- Wei, J.; Wang, X.; Schuurmans, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.
- Yan, B.; Zhou, Z.; Zhang, L.; et al. 2025. Beyond self-talk: A communication-centric survey of LLM-based multi-agent systems. *arXiv:2502.14321*.
- Yang, A.; Li, A.; Yang, B.; et al. 2025. Qwen3 technical report. *arXiv:2505.09388*.
- Zhao, H.; Zhao, C.; Li, Y.; Zhang, Z.; and Liu, G. 2025. Thinking in a Crowd: How Auxiliary Information Shapes LLM Reasoning. *arXiv:2509.18163*.
- Zhou, X.; Tie, G.; Zhang, G.; Wang, W.; Zuo, Z.; Wu, D.; Chu, D.; Zhou, P.; Gong, N. Z.; and Sun, L. 2025. Exploring the Necessity of Reasoning in LLM-based Agent Scenarios. *arXiv preprint arXiv:2503.11074*.
- Zhu, K.; Du, H.; Hong, Z.; et al. 2025a. Multiagentbench: Evaluating the collaboration and competition of LLM agents. *arXiv:2503.01935*.
- Zhu, X.; Zhang, C.; Stafford, T.; et al. 2025b. Conformity in Large Language Models. In *ACL*, 3854–3872. Austria.