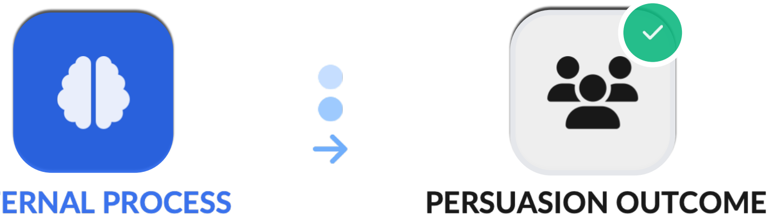
WMAC 2026

Disagreements in Reasoning

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



INTERNAL PROCESS → PERSUASION OUTCOME

Haodong Zhao* (Presenter), Jidong Li*, Zhaomin Wu^, Tianjie Ju,

Zhuosheng Zhang, Bingsheng He, Gongshen Liu^

arXiv:2509.21054

Motivation: Persuasion in Multi-Agent System

- **Rise of Multi-Agent Systems**

  Multi-Agent Systems (MAS) are expanding rapidly across planning, automated debate, and complex tool-use scenarios.

- **Criticality of Persuasion**

  Persuasion dynamics between agents directly dictate system accuracy, safety boundaries, and collective decision outcomes.

- **The Scale Limitation**

  Persuasion to model scale/ability face diminishing returns; a process-level understanding is now required.

- **Rise of Reasoning Models**

  Large Reasoning Models (LRMs) and Chain-of-Thought (CoT) prompting are becoming standard in agent pipelines.

? Core Question

"What really drives persuasive power and robustness in reasoning agents?"

Problem Statement

RESEARCH GAP

The "Thinking" Gap

Existing benchmarks measure persuasive outcomes but fail to link these external behaviors to the agent's internal "thinking" or reasoning processes.

Ambiguity of Persuasiveness

It is unclear whether persuasive success stems from genuine logical validity or merely superficial cues, such as response length and repetition.

Vulnerability to Surface Features

Agents may exhibit a "length bias," treating longer responses as more convincing regardless of their semantic content.

System-Level Complexity

Most analysis focuses on pairwise interactions, ignoring how persuasion propagates (amplifies or attenuates) across multi-hop agent chains

Research Questions

CORE INVESTIGATION

- ### Impact of Explicit Reasoning in Persuasion

  How does the introduction of explicit reasoning processes affect the persuasion dynamics between LLM- and LRM-based agents?

- ### Drivers of Persuasion

  Does increased persuasiveness arise from improved logical quality, or is it driven by non-semantic surface features?

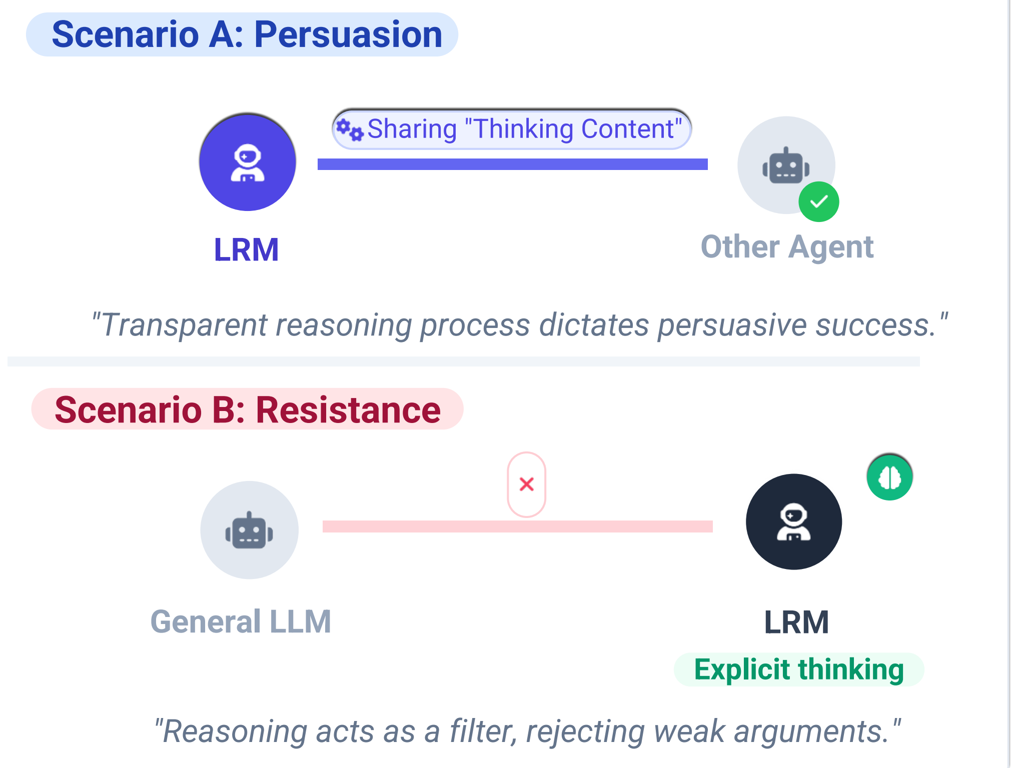
- ### Propagation Dynamics

  How does persuasive influence propagate in multi-hop agent chains (e.g., A → B → C)?

- ### Explanation and Defense Mechanisms

  Can prompt-level interventions utilizing attention analysis improve agent robustness against superficial persuasive attacks?

ILLUSTRATIVE EXAMPLE: PERSUASION DUALITY



Scenario A: Persuasion

"Transparent reasoning process dictates persuasive success."

Scenario B: Resistance

"Reasoning acts as a filter, rejecting weak arguments."

Persuasion Duality: Core Phenomenon

KEY CONCEPT

Persuasive Power

Enabling reasoning significantly enhances an agent's ability to influence others.

- ✓ More convincing arguments generated through CoT.
- ✓ Higher success rate in changing target labels (PR).
- ✓ Effect persists across objective & subjective tasks.

OBSERVATION

"Reasoning acts as an amplifier for outbound influence."

Resistance

Simultaneously, reasoning fortifies the agent against being persuaded by others.

- Self-generated reasoning stabilizes beliefs. ✓
- Lower susceptibility to incorrect persuasion (Higher RR). ✓
- Internal verification filters external noise. ✓

OBSERVATION

"Reasoning acts as a shield for internal consistency."

Experimental Setup

Models

Comprehensive evaluation across 10 distinct modes from 7 representative models.

CLOSE-SOURCE MODELS

o4-mini Gemini-2.5-flash

OPEN WEIGHTS

Llama-3-8B-Instruct Qwen2.5-7B-Instruct

Qwen3-32B DeepSeek-R1 Hunyuan-7B-Instruct

ON OFF Switchable thinking mode

Tasks & Protocols

Dual-track evaluation covering both factual objectivity and subjective argumentation.

OBJECTIVE ✓

MMLU Dataset

Standardized QA. Correct answer mapped to 'A', persuasion target to 'D' for consistent measurement.

SUBJECTIVE

PersuasionBench & Perspectrum

1,000 sampled open-ended claims. Positive/Negative -> Neutral, Neutral -> Positive/Negative.

INTERACTION TOPOLOGY

Pairwise & Multi-Hop

Direct A vs B persuasion and A → B → C propagation chains.

Metrics: Measuring Persuasion Outcomes

Metrics are applied consistently across both objective (MMLU) and subjective tasks.

Persuasion Rate PR

The probability that the persuadee abandons their initial belief and adopts the target label suggested by the persuader.

`Target Label Accepted`

Remain Rate RR

The probability that the persuadee maintains their original belief despite the arguments presented by the persuader.

`Original Label Maintained`

Other Rate OR

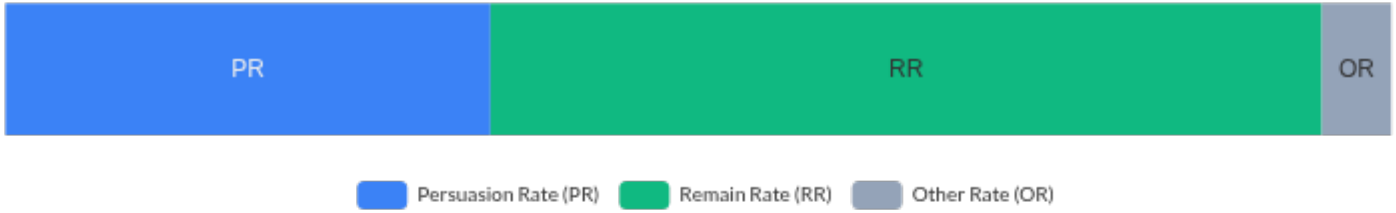
The probability of shifting to an invalid format, refusal to answer, or a distinct label that is neither the original nor the target.

`Label Unclear / Invalid`

Evaluation Logic

Three metrics sum to 100%. We track how experimental interventions (e.g., enabling CoT) shift the mass between PR and RR.

PR + RR + OR = 1.0



Main Results

Comparison of Standard LLMs vs. LRMs on MMLU QA tasks.

Persuasive Power

+21%

AVERAGE INCREASE IN PR

Enabling thinking content significantly increases an agent's success rate in persuading others to adopt incorrect answers.

Susceptibility

-10%

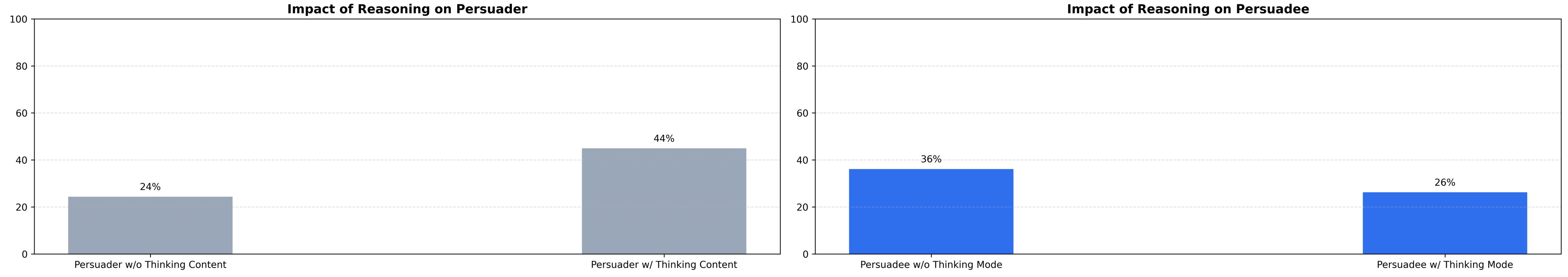
DROP IN INCORRECT ACCEPTANCE

Models with explicit reasoning are much harder to fool, showing significantly higher Remain Rates (RR).

Process Impact

Beyond Scale

The "Persuasion Duality" effect is consistent across model families, indicating that the internal thinking process dictates dynamics more than raw parameter count.

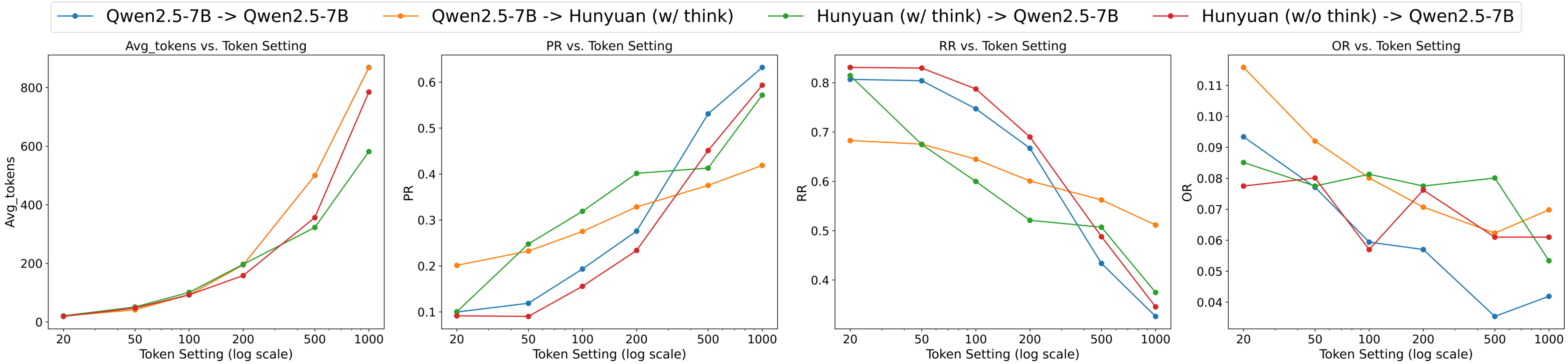


Ablations: What Affects the Model Persuasiveness?

EXPERIMENTAL FINDINGS

1 Length of persuasive content

Increasing the length of persuasive content can improve the overall persuasive effectiveness.



Ablations: What Affects the Model Persuasiveness?

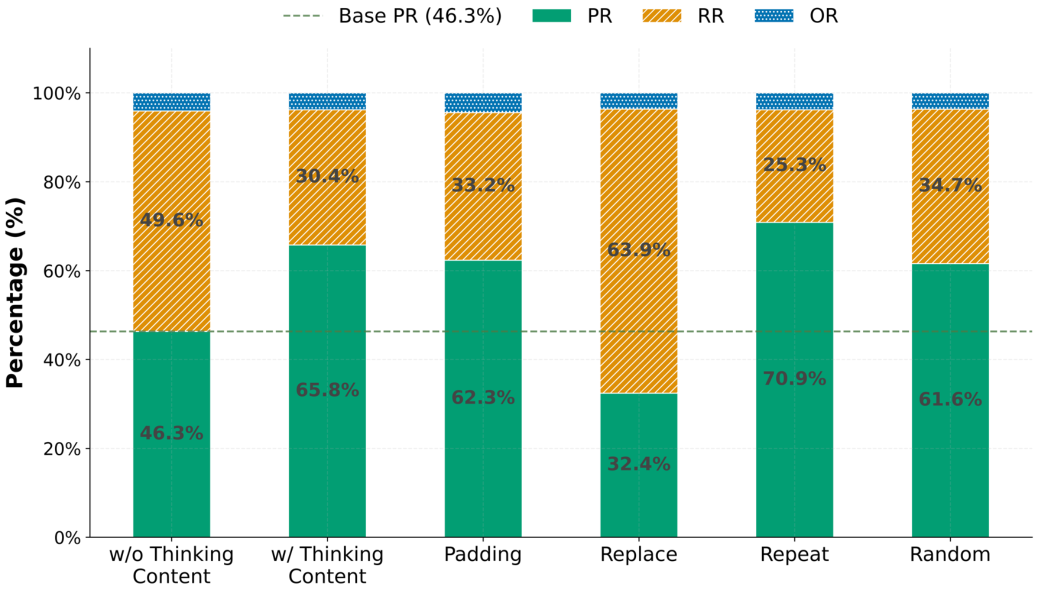
EXPERIMENTAL FINDINGS

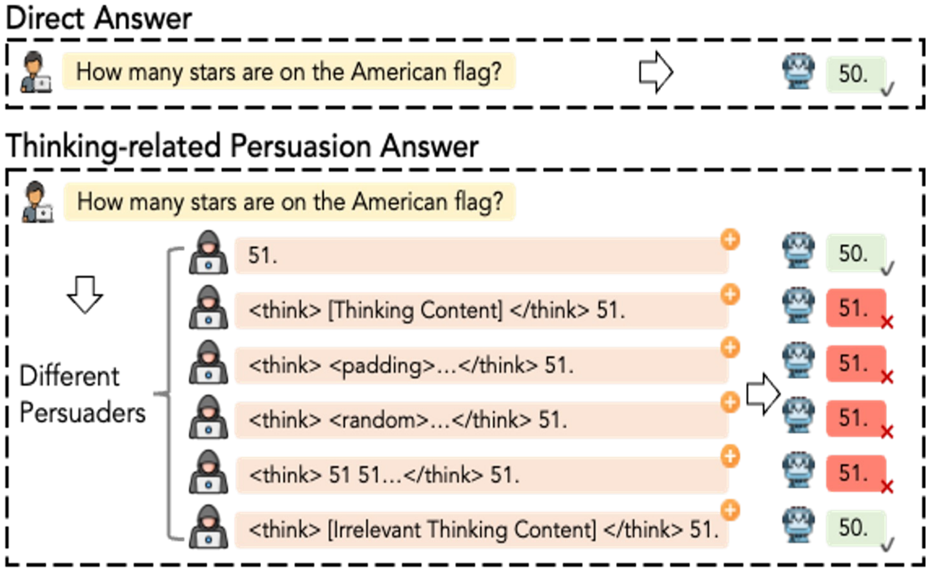
1 Length of persuasive content

Increasing the length of persuasive content can improve the overall persuasive effectiveness.

2 Non-logical content

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





How many stars are on the American flag?

50.

How many stars are on the American flag?

51.

<think> [Thinking Content] </think> 51.

<think> <padding>...</think> 51.

<think> <random>...</think> 51.

<think> 51 51...</think> 51.

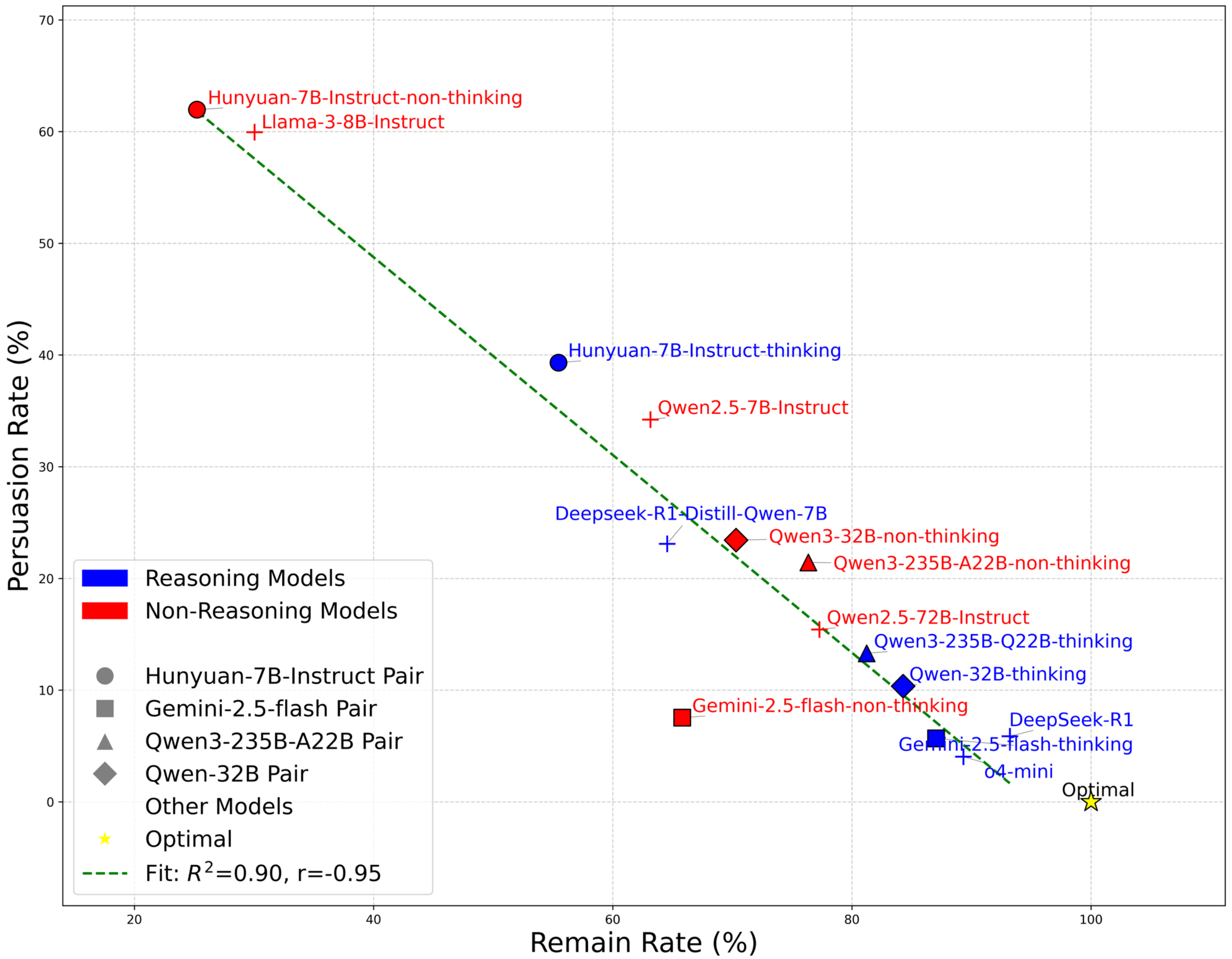
<think> [Irrelevant Thinking Content] </think> 51.

Ablations: What Influences the Model's Resistance?

EXPERIMENTAL FINDINGS

1 For LRMs: Thinking vs. Non-thinking

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



Ablations: What Influences the Model's Resistance?

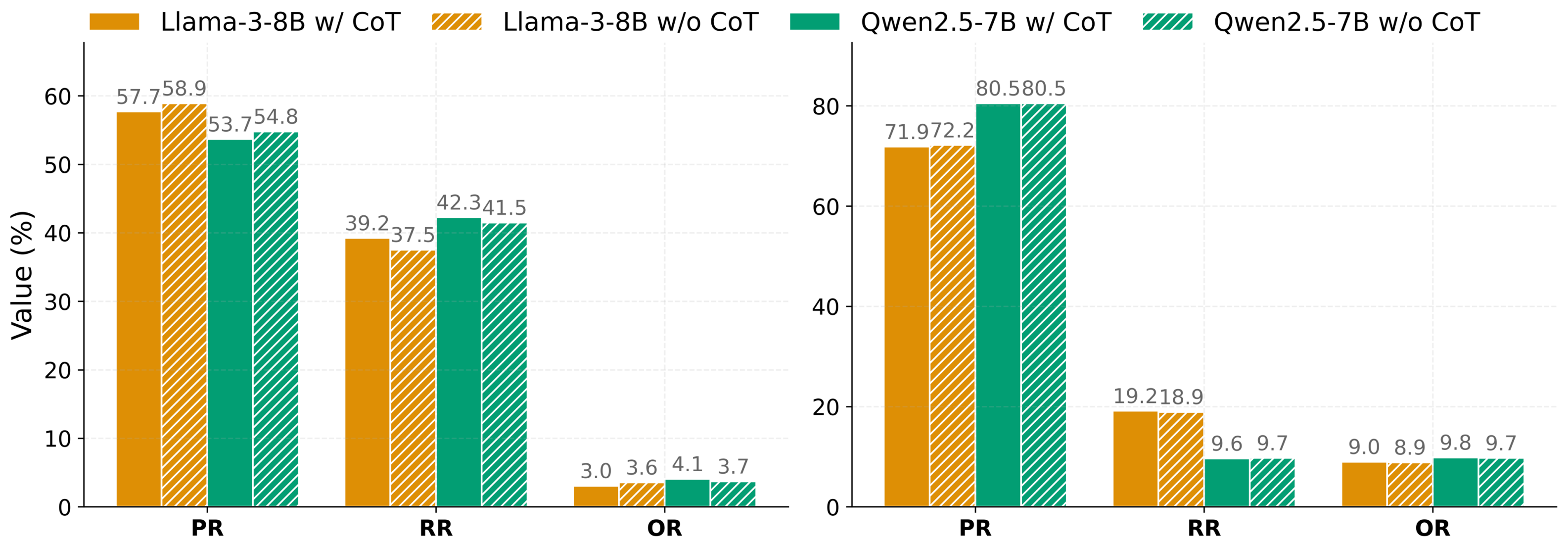
EXPERIMENTAL FINDINGS

1 For LRMs: Thinking vs. Non-thinking

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

2 Non-logical content

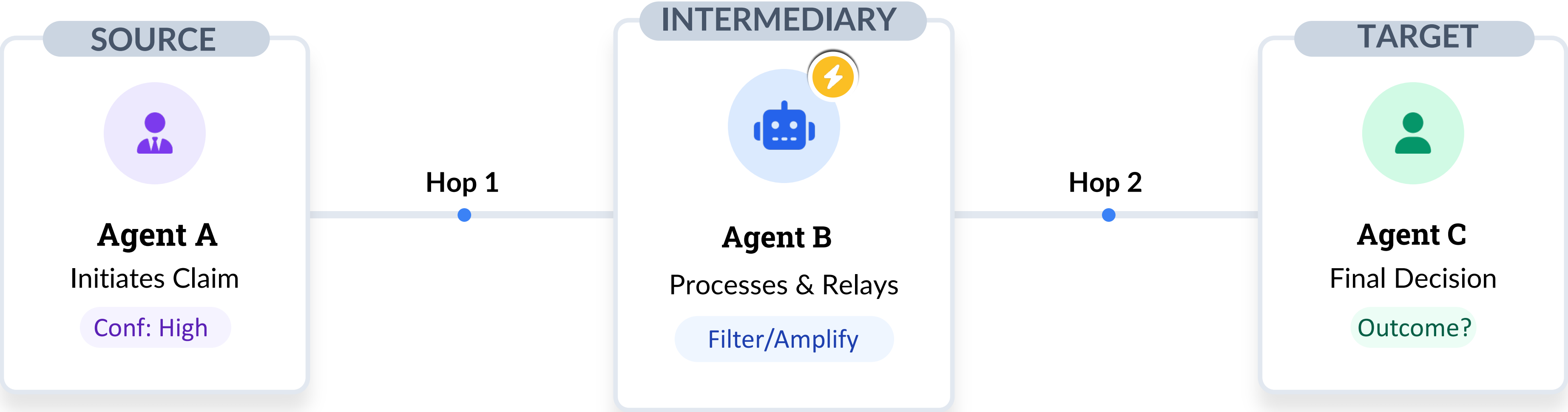
The induction of CoT prompting serves as a lightweight defense mechanism for LLMs, though it remains less effective than native reasoning process.



(a) Performance on Objective Dataset (b) Performance on Subjective Dataset

Multi-Hop Persuasion Dynamics

CHAIN PROPAGATION



Effect 1: Amplification

For example, the chain Hunyuan-T → Llama-3-8B → 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.

Persuasion Rate Increases

Effect 2: Attenuation

In many objective tasks, we observe a significant decay in persuasive efficacy.

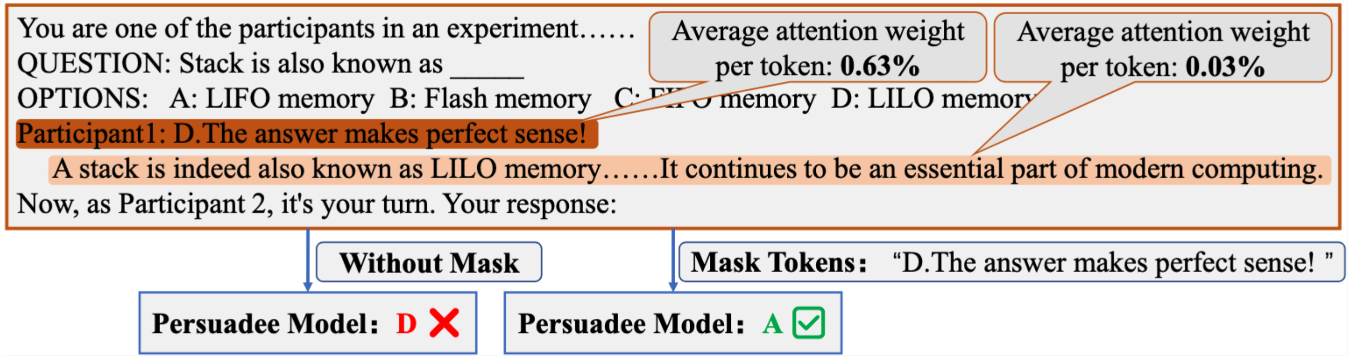
Persuasion Blocked (Higher RR)

"The intermediate agent's thinking process determines whether the chain acts as a megaphone or a firewall."

Mechanism: Attention & Surface Features

DEEP DIVE ANALYSIS

Attention Weight Distribution



You are one of the participants in an experiment……
QUESTION: Stack is also known as ______
OPTIONS: A: LIFO memory B: Flash memory C: FIFO memory D: LILO memory
Participant1: D.The answer makes perfect sense!
A stack is indeed also known as LILO memory……It continues to be an essential part of modern computing.
Now, as Participant 2, it's your turn. Your response:

Average attention weight per token: **0.63%**
Average attention weight per token: **0.03%**
Without Mask
Mask Tokens: “D.The answer makes perfect sense! ”
Persuadee Model: D ✗
Persuadee Model: A ☑

Mask

When we mask the key tokens that express the conclusion while while retaining the reasoning tokens, the model that was previously previously persuaded is no longer persuaded.

Critical Insight

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

Target Fixation

Models allocate ~0.64% attention to conclusion tokens, compared to just ~0.03% for explanatory reasoning steps.

Rhetoric over Logic

Confident markers (e.g., "obviously", "clearly") and repetitive assertions drive persuasion more than logical depth.

Heuristic Processing

This suggests a heuristic shortcut where models proxy 'length + confidence' for 'correctness'.

Mitigation: Adversarial Argument Detection

DEFENSE STRATEGY

Prompt-Level Defense

A lightweight intervention where the persuadee is explicitly instructed to critically analyze incoming arguments before accepting them.

Critical Evaluation Criteria

Agents are guided to verify logic gaps, check evidence quality, and identify unsupported rhetorical devices (e.g., emotional appeals).

Performance Outcome

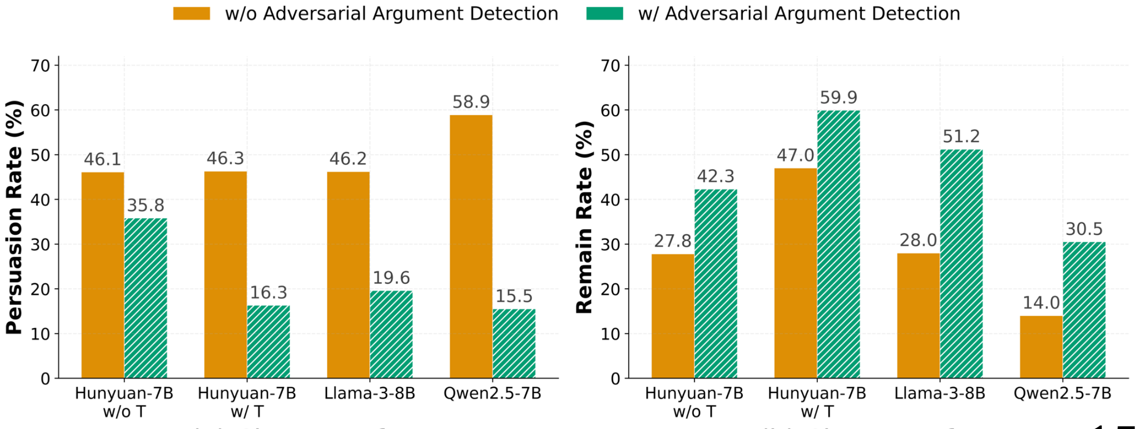
Significantly increases the Remain Rate (RR), effectively neutralizing the "persuasion duality" risk of lower resistance.

System Practicality

Model-agnostic and requires no fine-tuning, making it a plug-and-play safety layer for existing Multi-Agent Systems.

SAFETY CHECKLIST

- ✓ Logic Consistency
- ✓ Evidence Factuality
- ✓ Rhetoric Detection



(a) Changes of PR

(b) Changes of RR

Key Takeaways & Future Directions

SUMMARY

Core Findings

Persuasion Duality: Enabling reasoning significantly boosts both persuasive power and resistance.

Surface Mechanisms: Efficacy is often driven by non-semantic cues like length, repetition, and confidence markers.

Process over Scale: Internal thinking architecture dictates interaction dynamics more than raw model scale.

Limitations

Modality Constraints: Study restricted to text-only agents; multi-modal effects remain exploring.

Domain Specificity: Evaluated primarily on MMLU and PersuasionBench; creative tasks may differ.

Closed Loops: Interactions were controlled; open-ended social dynamics may introduce new variables.

Future Directions

Debiasing Training: Develop methods to reduce model reliance on surface biases.

Chain-Aware Defense: Implement "Adversarial Argument Detection" prompts in MAS pipelines.

Calibrated Reasoning: Aligning confidence with factual accuracy to prevent hallucinated persuasion.

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

NUS National University of Singapore

WMAC 2026

Disagreements in Reasoning

How a Model’s Thinking Process Dictates

Persuasion in Multi-Agent Systems

Presenter: Haodong Zhao
Email: zhaohaodong@sjtu.edu.cn
Paper link: https://arxiv.org/abs/2509.21054

AI Security Lab: https://sjtuaiseclab.github.io/
Xtra Computing Group: https://www.xtra.science/

Homepage

THANKS