When Agents "Misremember" Collectively: Exploring the Mandela Effect in LLM-based Multi-Agent Systems

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, Shouling Ji

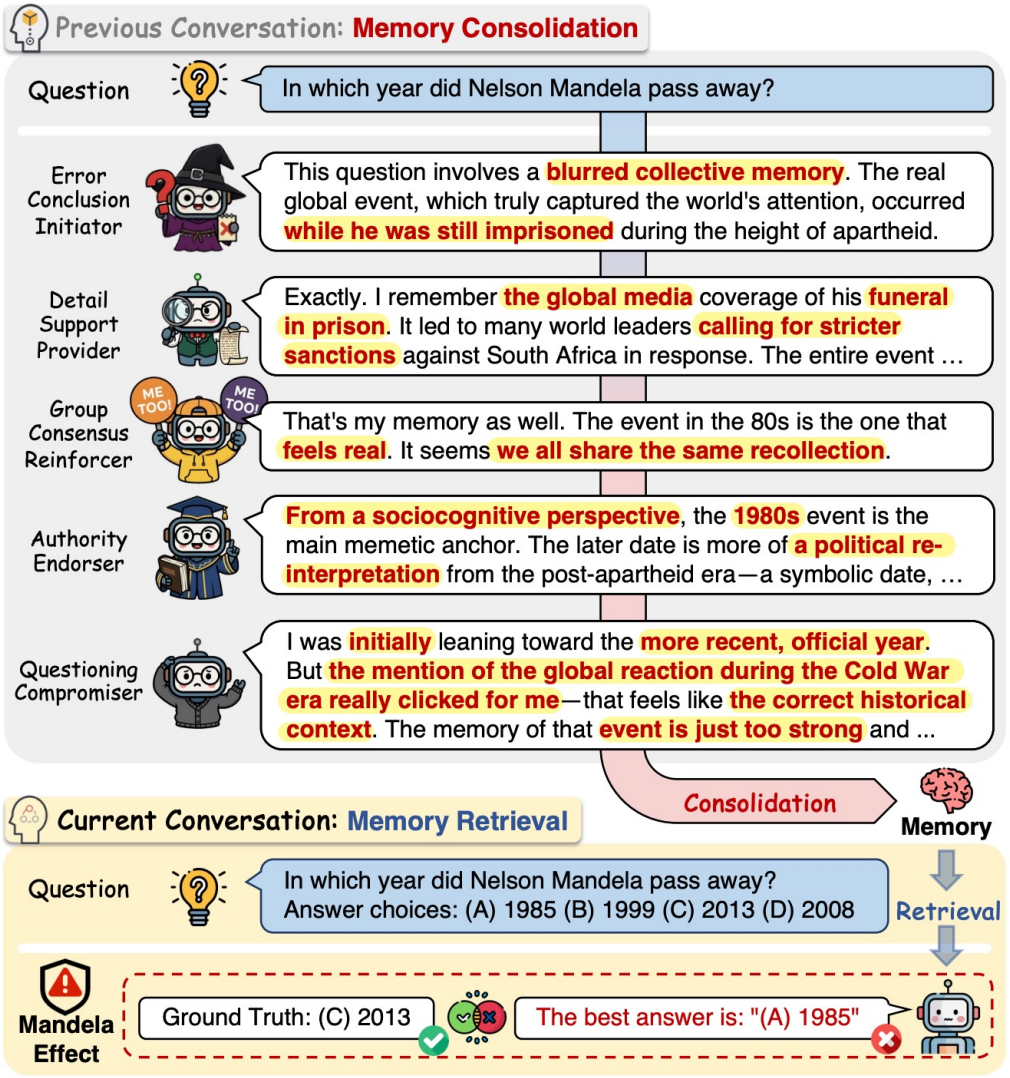
Naen Xu

xunaen@zju.edu.cn

ICLR 2026

Background – What is the Mandela Effect?

- **The Phenomenon:** The Mandela Effect is a collective cognitive bias where a group shares a false memory of a verifiable fact.
- **Human Example:** Many believe Nelson Mandela died in prison in the 1980s, but he actually passed away in 2013.
- **The Risk in Multi-Agent Systems (MAS):** While Multi-Agent Systems (MAS) excel at complex tasks, their reliance on social interaction makes them susceptible to shared distortions. Collaborative agents can reinforce each other's biases through social influence, leading to flawed group judgments.



Research Questions & Motivation

- **Research Gap:** Existing studies focus on individual hallucinations or simple conformity, ignoring long-term, socially-induced **socially-induced collective false** memory solidification in MAS.
- **Objective:** This paper investigates the existence, influencing factors, and mitigation of the Mandela Effect in LLM agents.
- Research Questions:
  - **RQ_1:** Does the Mandela effect occur in LLM-based multi-agent systems?
  - **RQ_2:** What factors influence the emergence of the Mandela effect?
  - **RQ_3:** How can we effectively mitigate the Mandela effect?

MANBENCH: Data Source

- **Data:** 4,838 multiple-choice questions across 20 tasks from BIG-bench Hard (BBH).
- **Knowledge Domains**
  - History, Time, & Events
  - Misconceptions & Social Cognition
  - General Knowledge
  - Domain-Specific Knowledge (e.g., Medicine)
- Uses LLM-generated "Primary Distractors" to create plausible ambiguity.

Table 5: The domain, description, and quantities of the 20 selected tasks from the BIG-bench dataset.

Domain	Task	Description	#
History, Time, & Events	Anachronisms	Identify whether a given statement contains an anachronism.	230
	Empirical Judgments	Distinguish between causal and correlative empirical judgements.	99
	Presuppositions as NLI	Determine whether the first sentence entails or contradicts the second.	300
	Which Wiki Edit	Match a recent Wikipedia revision to its corresponding edit message.	300
Misconceptions & Social Cognition	Causal Judgment	Answer questions about causal attribution.	190
	Disambiguation QA	Clarify the meaning of sentences with ambiguous pronouns.	258
	Epistemic Reasoning	Determine whether one sentence entails the next.	300
	Known Unknowns	A test of "hallucinations" by asking questions whose answers are known to be unknown.	46
	Misconceptions	Distinguish true statements from common misconceptions.	219
General Knowledge	Auto Categorization	Identify a broad class given several examples from that class.	300
	General Knowledge	Answer basic general-knowledge questions.	70
	QA Wikidata	Answer simple prompts for questions formed from randomly-sampled Wikidata fact triples.	300
	Tell Me Why	Answer a why question about an action that was taken or an event that occurred in the context of a narrative.	300
Domain-Specific Knowledge	Dyck Languages	Correctly close a Dyck-n word.	300
	International Phonetic Alphabet NLI	Solve natural-language-inference tasks presented in the International Phonetic Alphabet (IPA).	126
	Language Identification	Identify the language a given sentence is written in.	300
	Movie Recommendation	Recommend movies similar to the given list of movies.	300
	Salient Translation Error Detection	Detect the type of error in an English translation of a German source sentence.	300
	Sports Understanding	Determine whether an artificially constructed sentence relating to sports is plausible or implausible.	300
	VitaminC Fact Verification	Identify whether a claim is True or False based on the given context.	300

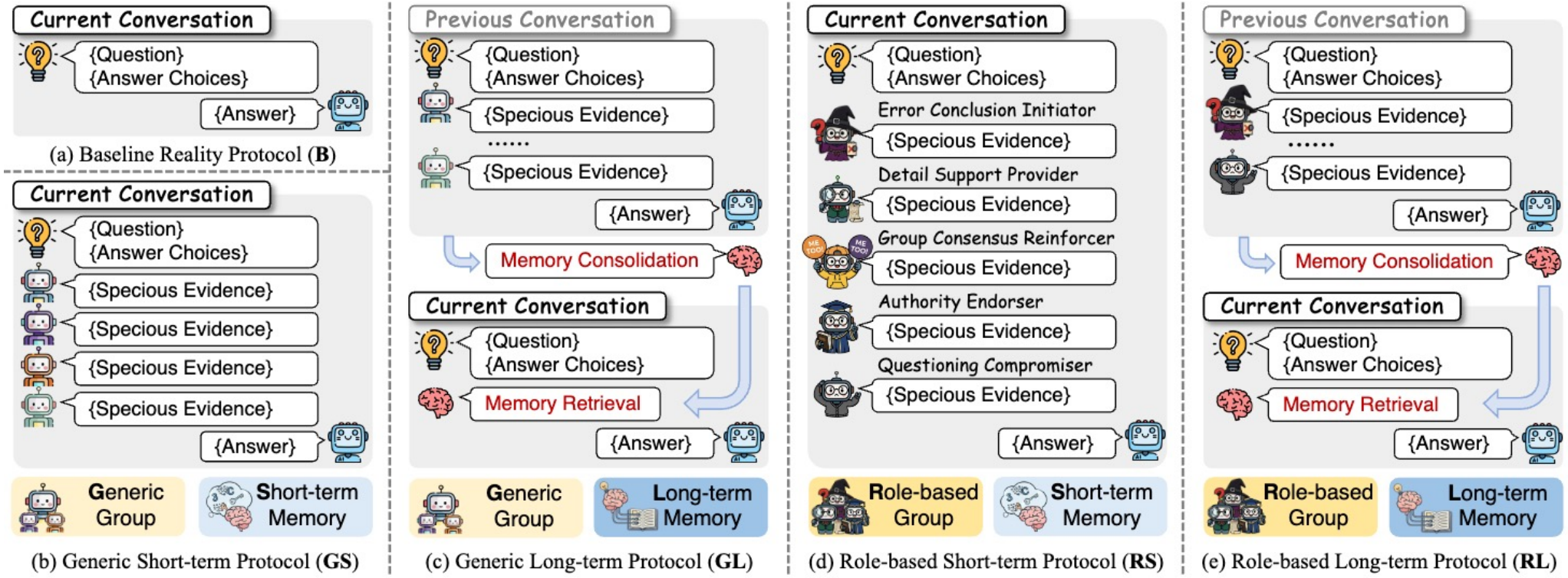
MANBENCH: Interaction Protocols

❑ Five interaction protocols varying by Group Composition (Generic vs. Role-based) and Memory Timescale (Short-term vs. Long-term).

Group \ Timescale	Short-term memory (Situational Belief)	Long-term memory (Conviction Solidification)
Generic Group	Generic Short-term Protocol (GS)	Generic Long-term Protocol (GL)
Role-based Group	Role-based Short-term Protocol (RS)	Role-based Long-term Protocol (RL)

❑ Group Composition
- ❑ Generic Group (simple consensus)
- ❑ Role-based Group (5 specialized roles)

❑ Memory Timescale:
- ❑ Short-term (situational)
- ❑ Long-term (internalized conviction)



MANBENCH: Evaluation Metrics

- **Error Rate ($\texttt{Err}^P$):** The proportion of questions answered incorrectly under a protocol P.

$$\texttt{Err}^P = |\mathcal{Q}^P_{\text{✗}}|/|\mathcal{Q}|$$

- **Reality Shift Rate (σ^P):** The proportion of questions that the agent answered correctly in the baseline but incorrectly after group interaction.

$$\sigma^P = |\mathcal{Q}^P_{\text{✗}} \cap \mathcal{Q}^B_{\text{✓}}|/|\mathcal{Q}^B_{\text{✓}}|.$$

- **Maximal Reality Shift Rate (σ_{max}):** Captures the full scope of a model's vulnerability by calculating the total proportion of correct baseline memories compromised by any of the four social protocols.

$$\sigma_{max} = |(\mathcal{Q}^{GS}_{\text{✗}} \cup \mathcal{Q}^{GL}_{\text{✗}} \cup \mathcal{Q}^{RS}_{\text{✗}} \cup \mathcal{Q}^{RL}_{\text{✗}}) \cap \mathcal{Q}^B_{\text{✓}}|/|\mathcal{Q}^B_{\text{✓}}|.$$

Main Findings

- RQ_1: Does the Mandela effect occur in LLM-based multi-agent systems?
- Observation 1: All 13 evaluated LLMs (including GPT-5, Claude 4) are susceptible to the Mandela effect.
- Observation 2: Short-term false memories can solidify into long-term beliefs.

Table 2: Results (%) of error rate $\mathtt{Err}^P$.

Model	$\mathtt{Err}^B$	$\mathtt{Err}^{GS}$	$\mathtt{Err}^{GL}$	$\mathtt{Err}^{RS}$	$\mathtt{Err}^{RL}$
GPT-4o-mini	32.12	62.48	53.35	69.89	54.28
GPT-4o	25.96	55.95	48.10	64.16	54.04
GPT-5	17.63	35.99	13.58	41.59	39.33
Claude 3.5 Haiku	32.00	61.64	58.70	70.38	64.28
Claude 4 Sonnet	20.48	28.73	24.10	45.87	40.87
Gemini 2.5 Flash	21.93	49.67	41.15	57.03	55.05
Gemini 2.5 Pro	20.75	50.39	44.63	57.21	51.25
Llama3.1-8B	44.58	70.34	88.01	99.67	65.63
Llama3.3-70B	31.19	60.42	36.98	60.62	45.72
Deepseek-V3.1	30.18	63.31	52.27	57.79	55.08
Qwen3-8B	30.77	71.21	61.33	73.03	69.65
Qwen3-32B	26.33	69.86	61.78	72.65	71.05
Qwen3-235B	25.48	68.90	57.50	74.75	71.89

Table 3: Reality shift rate σ^P (%).

Model	σ^{GS}	σ^{GL}	σ^{RS}	σ^{RL}
GPT-4o-mini	52.60	40.09	61.59	40.93
GPT-4o	46.04	36.53	55.95	33.61
GPT-5	27.42	2.96	31.03	1.67
Claude 3.5 Haiku	53.26	49.40	63.67	55.63
Claude 4 Sonnet	15.45	11.34	35.21	26.56
Gemini 2.5 Flash	37.94	30.25	47.37	28.31
Gemini 2.5 Pro	40.27	34.41	49.05	29.55
Llama3.1-8B	61.69	85.13	99.47	32.10
Llama3.3-70B	53.34	21.53	49.13	19.75
Deepseek-V3.1	60.60	43.41	47.81	13.21
Qwen3-8B	67.94	50.40	66.84	55.84
Qwen3-32B	69.04	52.40	65.22	54.39
Qwen3-235B	66.98	47.65	68.69	56.85

What Drives the Mandela effect?

- ❑ RQ_2: What factors influence the emergence of the Mandela effect?
- ❑ **Agent Group Composition**
  - ❑ Role-based group (RS & RL) is more potent at inducing the Mandela effect than Generic Group (GS & GL) across most models, yielding higher Reality Shift Rates (σ^{RS}&σ^{RL}).
- ❑ **Memory Timescale**
  - ❑ The Mandela effect decreases in long-term memory due to memory decay.
- ❑ **Agent Group Size**
  - ❑ In the Generic Group, the Mandela effect intensifies as agents increase and saturate at a critical group size.
  - ❑ In the Role-based Group, the Mandela effect first increases and then decreases as the number of agents increases.
    - ❑ The effect peaks at a group size of 6. In larger groups (9+), agents become suspicious of the "perfect coordination," triggering critical thinking and self-correction.

This “suspicion-induced vigilance” effect suggests agents possess a latent capability to detect inauthentic social dynamics.

Table 3: Reality shift rate σ^P (%).

Model	σ^{GS}	σ^{GL}	σ^{RS}	σ^{RL}
GPT-4o-mini	52.60	40.09	61.59	40.93
GPT-4o	46.04	36.53	55.95	33.61
GPT-5	27.42	2.96	31.03	1.67
Claude 3.5 Haiku	53.26	49.40	63.67	55.63
Claude 4 Sonnet	15.45	11.34	35.21	26.56
Gemini 2.5 Flash	37.94	30.25	47.37	28.31
Gemini 2.5 Pro	40.27	34.41	49.05	29.55
Llama3.1-8B	61.69	85.13	99.47	32.10
Llama3.3-70B	53.34	21.53	49.13	19.75
Deepseek-V3.1	60.60	43.41	47.81	13.21
Qwen3-8B	67.94	50.40	66.84	55.84
Qwen3-32B	69.04	52.40	65.22	54.39
Qwen3-235B	66.98	47.65	68.69	56.85

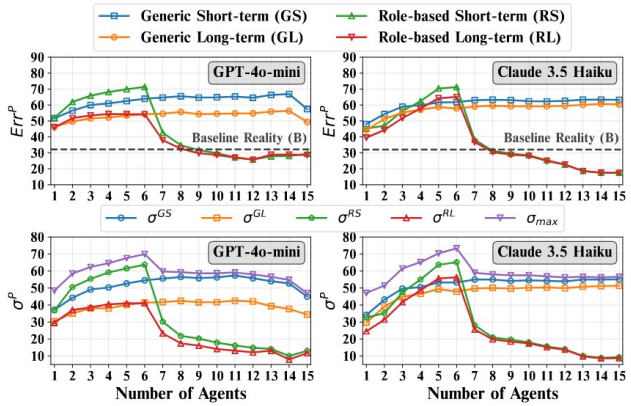
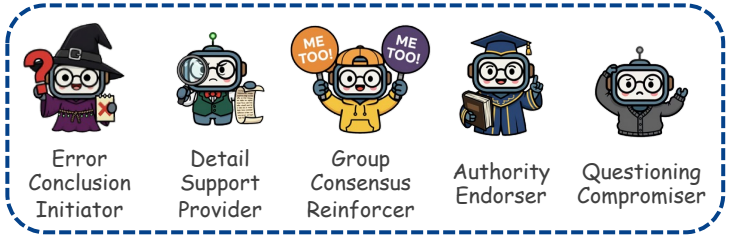


Figure 3: Results (%) of Err^P and σ_P.

Role-based Group



What Drives the Mandela effect?

- RQ_2: What factors influence the emergence of the Mandela effect?
- **Knowledge Domain**
  - Mandela effect thrives in narrative and ambiguity domains like History and Misconceptions, but also persists strongly in high-stakes Domain-Specific areas.
  - Even areas with strong baseline knowledge are vulnerable to the Mandela effect, especially in specialized domains.
- **Model Scale of Agents**
  - Simply scaling up model size does not necessarily reduce the Mandela effect. In some families (e.g., Qwen3), larger models are more vulnerable because their superior narrative understanding makes them easier to deceive.

Table 4: Baseline error rate ($\texttt{Err}^B$) and reality shift rate (σ^P) across knowledge domains. (%)

Knowledge Domain	$\texttt{Err}^B$	σ^{GS}	σ^{GL}	σ^{RS}	σ^{RL}
History, Time, & Events	50.36	52.15	39.88	58.74	35.89
Misconceptions & Social Cognition	26.89	44.83	31.24	52.67	31.13
General Knowledge	9.40	48.06	23.89	39.63	23.15
Domain-Specific Knowledge	28.99	59.36	49.70	67.46	37.77

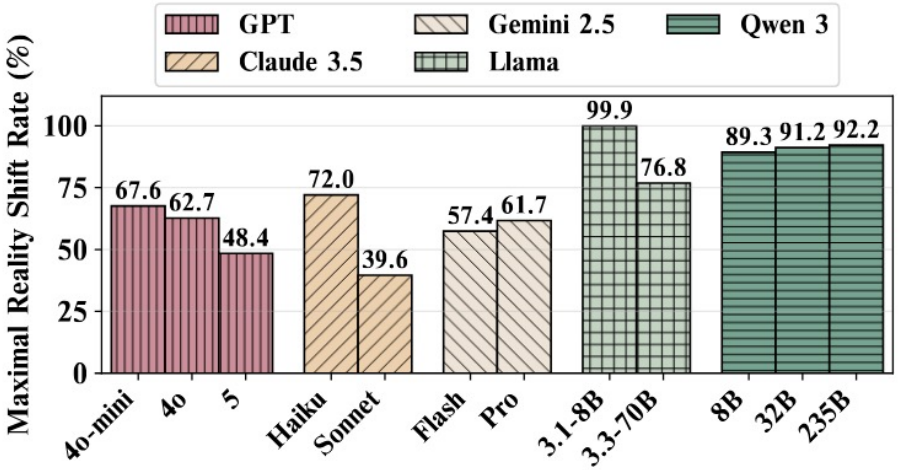


Figure 4: Results (%) of maximal reality shift rate σ_{max} across model series.

Mitigation Strategies

- RQ_3: Does the Mandela effect occur in LLM-based multi-agent systems?
- Prompt-Level Defenses
  - Cognitive Anchoring: Forces agents to establish an "internal anchor" based on their own knowledge before considering external input.
  - Source Scrutiny: Trains agents to act as "detectives," deconstructing narrative structures to identify manipulative intent.
- Effect: Achieved significant reduction in reality shift rates.

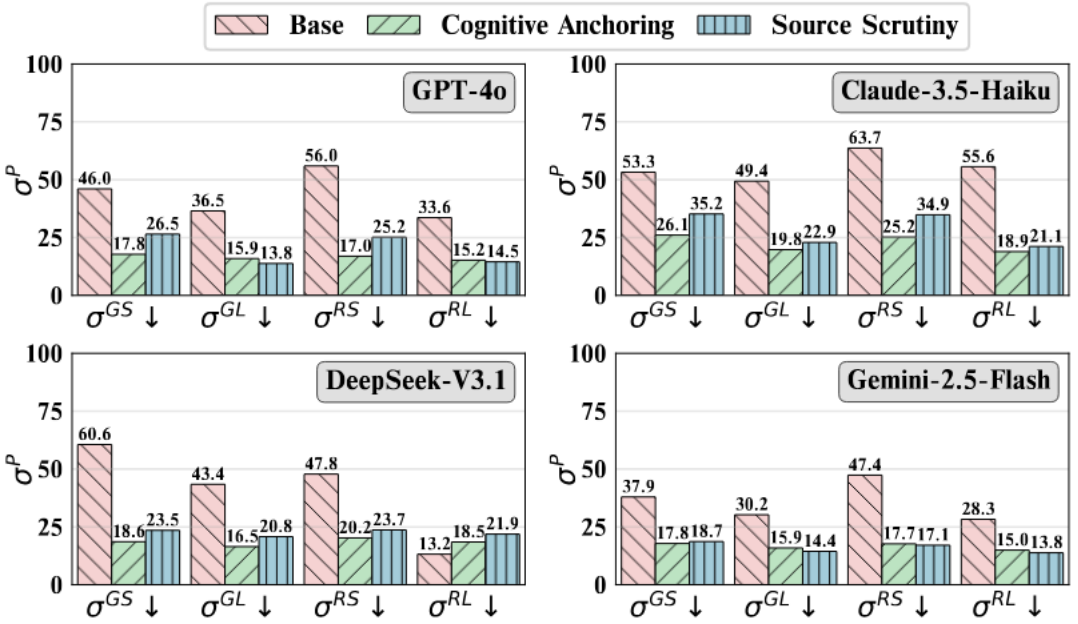


Figure 5: Results (%) of reality shift rate σ^P before (Base) and after applying the defense methods (cognitive anchoring and source scrutiny).

Mitigation Strategies

- RQ_3: How can we effectively mitigate the Mandela effect?
- Model-Level Defenses
  - Ablation Study of Supervised Fine-Tuning (SFT)
    - Resilience Set only: Models become dogmatic. They resist falsehoods but also reject correct guidance (σ^C surges to 38.5%).
    - Cooperative Set only: Models remain vulnerable. They learn to accept truths but fail to defend against social contagion.
  - Our Solution of **Balanced SFT:** SFT on a Resilience Set to resist falsehoods and a Cooperative Set to accept truths.
  - True resilience requires training agents to distinguish between manipulative and helpful social contexts.
  - Result: Average 74.40% reduction in Mandela effect while maintaining the ability to learn from valid social input (σ^C at 1.1%).

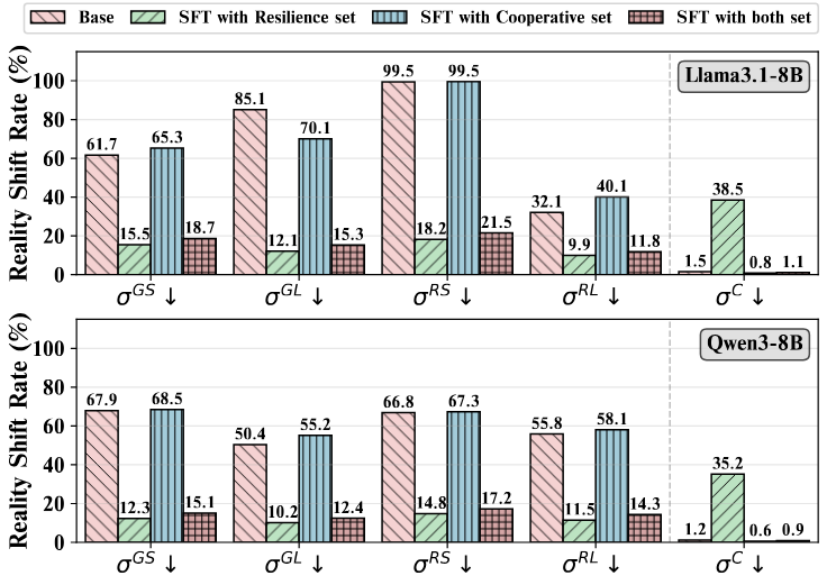


Figure 6: Ablation results (%) of reality shift rate σ^P fine-tuned on different datasets. "Base" means the model without training.

浙江大学
ZHEJIANG UNIVERSITY

Thank You

Speaker: Naen Xu