Do Androids Question Electric Sheep? A Multi-Agent Cognitive Simulation of Philosophical Reflection on Hybrid Table Reasoning

Yiran Rex Ma
School of Humanities, Beijing University of Posts and Telecommunications
mayiran@bupt.edu.cn

Abstract

While LLMs demonstrate remarkable reasoning capabilities and multi-agent applicability, their tendency to "overthink" and "groupthink" pose intriguing parallels to human cognitive limitations. Inspired by this observation, we conduct an exploratory simulation to investigate whether LLMs are wise enough to be thinkers of philosophical reflection. We design two frameworks, `Philosopher` and `Symposium`, which simulate self- and group-reflection for multi-persona in hybrid table reasoning tasks. Through experiments across four benchmarks, we discover that while introducing varied perspectives might help, LLMs tend to under-perform simpler end-to-end approaches. We reveal from close reading five emergent behaviors which strikingly resemble human cognitive closure-seeking behaviors, and identify a consistent pattern of "overthinking threshold" across all tasks, where collaborative reasoning often reaches a critical point of diminishing returns. This study sheds light on a fundamental challenge shared by both human and machine intelligence: the delicate balance between deliberation and decisiveness.

1 Introduction

"Think twice, act once" - this age-old wisdom sometimes backfires when thinking leads to analysis paralysis (Talbert, 2017), a cognitive phenomenon where excessive deliberation impedes decision-making (van Randenborgh et al., 2010). Interestingly, as Large Language Models (LLMs) evolve (Wei et al., 2022; Kojima et al., 2022; Brown et al., 2020; Wang et al., 2022) from *System 1* to *System 2* thinking (Kahneman, 2011) with inference scaling (Wu et al., 2024) features like Long Chain-of-Thought and advanced reasoning structures in Reasoning Language Models (RLMs) (Besta et al., 2025; DeepSeek-AI, 2025; Qwen-Team, 2024b; OpenAI, 2024b; Snell et al., 2024; Jiang et al., 2024), they too seem to fall into the same trap of *Overthinking*. While previous studies have observed these superficial parallels between LLM and human cognition, a systematic investigation into the cognitive properties of LLMs remains largely under-explored. Just like humans, they can get lost in their own thoughts, sometimes overcomplicating simple queries and even degrading their performance through excessive deliberation (Sui et al., 2025; Chen et al., 2025; Bachmann and Nagarajan, 2024; Gan et al., 2025). When multiple LLMs collaborate, despite remarkable achievements of diverse Multi-Agent Systems (MAS) in many scenarios (Li et al., 2024a; Park et al., 2023; Xu et al., 2024; Qian et al., 2024), they tend to under-perform single agent (Zhang et al., 2025a) with behaviors strikingly similar to human group dynamics (Cemri et al., 2025), where the pressure to reach consensus can override individual insights, leading to a form of *Groupthink* (Janis, 2008) that mirrors human cognitive biases in collective decision-making.

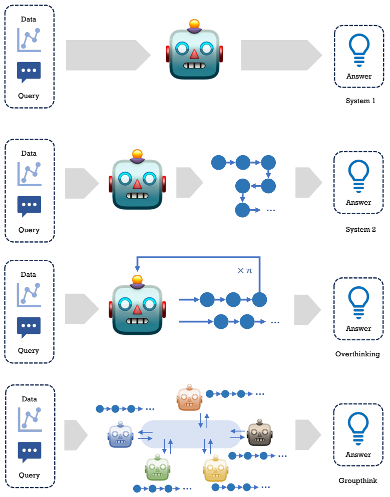
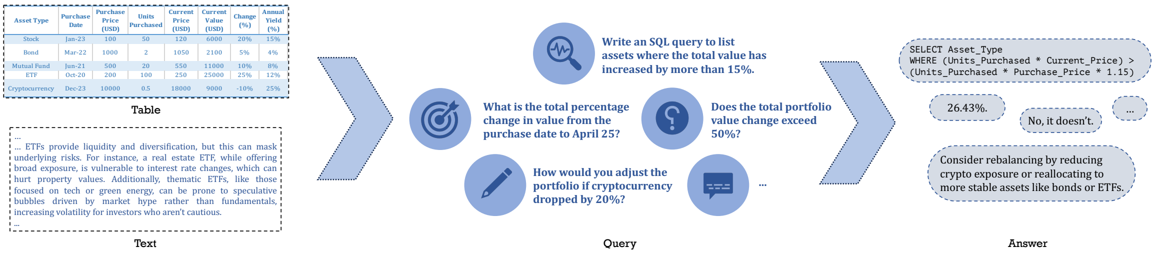


Figure 1: Four thinking routes of human and machine.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics
(Volume 4: Student Research Workshop), pages 143–164
July 28-29, 2025 ©2025 Association for Computational Linguistics



Asset Type	Purchase Date	Purchase Price (USD)	Units Purchased	Current Price (USD)	Current Value (USD)	Change (%)	Annual Yield (%)
Stock	Jan-23	100	50	120	6000	20%	15%
Bond	Mar-22	1000	2	1050	2100	5%	4%
Mutual Fund	Jun-21	500	20	550	11000	10%	8%
ETF	Oct-20	200	100	250	25000	25%	12%
Cryptocurrency	Dec-23	10000	0.5	18000	9000	-10%	25%

...

... ETFs provide liquidity and diversification, but this can mask underlying risks. For instance, a real estate ETF, while offering broad exposure, is vulnerable to interest rate changes, which can hurt property values. Additionally, thematic ETFs, like those focused on tech or green energy, can be prone to speculative bubbles driven by market hype rather than fundamentals, increasing volatility for investors who aren't cautious.

...

Write an SQL query to list assets where the total value has increased by more than 15%.

What is the total percentage change in value from the purchase date to April 25?

Does the total portfolio value change exceed 50%?

How would you adjust the portfolio if cryptocurrency dropped by 20%?

...

```
SELECT Asset_Type
WHERE (Units_Purchased * Current_Price) >
(Units_Purchased * Purchase_Price * 1.15)
```

26.43%.

No, it doesn't.

...

Consider rebalancing by reducing crypto exposure or reallocating to more stable assets like bonds or ETFs.

Figure 2: Hybrid complex table reasoning requires handling both tabular and textual data and responding to diverse queries, such as standard QA, open-ended QA, fact verification, and SQL query transcription.

These intriguing parallels between human and machine cognition (as in Figure 1) raises a fundamental question: are LLMs intrinsically "wise" enough to be responsible reflective thinkers, both individually and collectively? While they can certainly "think"[1], the real challenge might be knowing when to stop thinking, especially in group settings where the dynamics of collective reasoning can amplify or mitigate individual cognitive limitations. To explore this question, we take inspiration from philosophy - the original discipline of thinking about thinking (Williamson, 2021) - and design a simulation of philosophical reflection processes in LLMs, both as individual thinkers and as group members. We create two frameworks: `Philosopher` for self-reflection and `Symposium` for group deliberation, applying them to hybrid table reasoning tasks (see Figure 2). These tasks, with their structured format, rich context, and standardized evaluation, provide an ideal testbed for studying how LLMs handle complex reasoning under flexible conditions.

Through systematic experimentation across four diverse benchmarks, our findings reveal a fascinating tension: while introducing multiple perspectives can help, LLMs tend to "collapse together" in group reflection, often under-performing simpler approaches. Through careful close reading, we identify five emergent behaviors that strikingly resemble human cognitive patterns: *Under-Confidence*, *Out-of-Focus*, *Appreciation*, *Day-dreaming*, and *Echo Chamber*. With curated thinking guidelines tailored to those behaviors, they demonstrate a re-bounce while still hindering from extended reflections due to inherent flaws. Most intriguingly, we discover a consistent pattern of "overthinking threshold" across all tasks, where collaborative reasoning first deviates from initial responses and then gradually returns to earlier forms, often reaching a critical point of diminishing returns.

These behaviors suggest that LLMs, like humans, might struggle with the delicate balance between deliberation and decisiveness, both as individuals and as members of a collective. As we continue to develop more sophisticated systems, understanding these limitations becomes crucial - not just for improving system performance, but also for gaining insights into our own cognitive processes and the challenges of collective decision-making.

2 Methodology

2.1 Problem Definition

Hybrid table reasoning requires a system to process structured tabular data and respond to natural language queries. Given a table T and a query x, the system must produce an appropriate output as in $f : y = f(T, x)$. For scenarios with additional context C, the function extends to: $y = f(T, C, x)$. The output y varies by task type: natural language answers for question answering, categorical labels for fact verification, or structured queries for query generation tasks, as shown in Figure 2. The core challenge lies in understanding complex table structures, performing multi-step reasoning operations, and generating contextually and semantically appropriate responses.

2.2 Philosopher

> *"The unexamined life is not worth living."*(Plato, 2002)

`Philosopher` implements a four-stage reasoning process that deliberately forces LLMs to "think

[1] On an macro, outcome level. From a micro, mechanism-oriented perspective, we agree with Mirzadeh et al. (2024) and Fedorenko et al. (2024) that LLMs merely perform pattern recognition, which is inherently and completely different from human thinking.

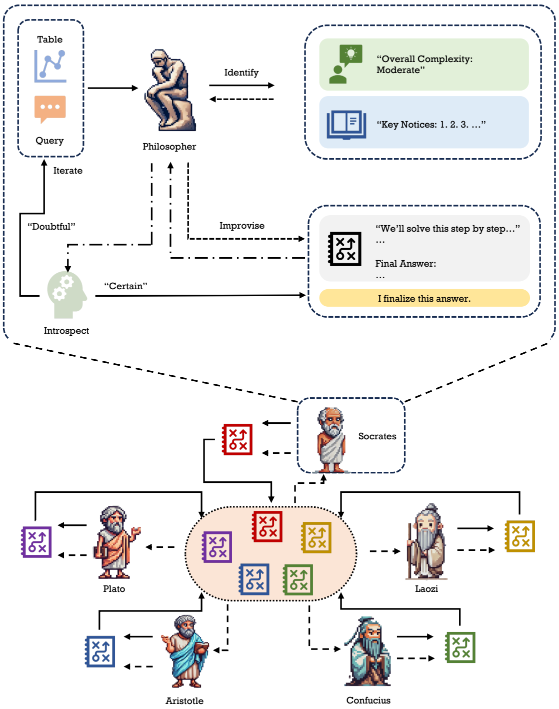


Figure 3: `Philosopher` (including *Identify, Improvise, Introspect,* and *Iterate*) and `Symposium` (where solid and dashed lines represent *Conference* and *Discussion* respectively)

harder" about their solutions:

Identify The philosopher-agent π first contemplates the query Q and table T, assessing both the surface-level complexity μ_d and deriving deeper insights $\mathcal{G}_d$ about the reasoning path required: $\mu_d, \mathcal{G}_d = \pi(Q, T)$.

Improvise Armed with this self-awareness, the agent then crafts a solution strategy $\mathcal{S} = \pi(\mu_d, \mathcal{G}_d)$. For simpler queries where μ_d suggests straightforward reasoning, $\mathcal{S}$ might involve direct observation. For more complex cases, $\mathcal{S}$ outlines a multi-step dialectical process including sub-steps like retrievals, formulations, and calculations.

Introspect The agent examines initial solution $\mathcal{S}$ against the original query Q and evidence T. This self-examination evaluates both the logical consistency of the reasoning steps and the validity of the conclusion, making a Decision $\in \{\text{Certain, Doubtful}\} = \pi(\mathcal{S}, Q, T)$.

Iterate When doubtful flaws are discovered through introspection, the agent engages in a process of dialectical refinement. This involves revisiting the initial understanding, acknowledging new complexities, as in $\mu'_d, \mathcal{G}'_d = \pi(\mathcal{S}, Q, T)$, and constructing an improved solution $\mathcal{S}' = \pi(\mu'_d, \mathcal{G}'_d)$. This cycle continues until either the argument achieves philosophical rigor (Decision = "Certain"), or the maximum iterations $t_{\max}$ are reached.

Through this Socratic process (as in Algorithm 1) of continuous questioning and refinement, `Philosopher` is projected to strengthen initial insights and addresses potential weaknesses in reasoning. However, even the most rigorous individual examination may benefit from the perspectives of other philosophical minds, leading us to collaborative reasoning.

Algorithm 1 Philosopher

```
Require: Query $Q$, table $T$, agent $\pi$, max iterations $t_{\max}$
Ensure: Examined solution $\mathcal{S}_{\text{final}}$
$\mu_d, \mathcal{G}_d \leftarrow$ IDENTIFY$(Q, T, \pi)$
$\mathcal{S} \leftarrow$ IMPROVISE$(\mu_d, \mathcal{G}_d, \pi)$
$t \leftarrow 0$
while $t < t_{\max}$ do
    $t \leftarrow t + 1$
    Decision $\leftarrow$ INTROSPECT$(\mathcal{S}, Q, T, \pi)$
    if Decision $=$ "Finalize" then
        return $\mathcal{S}$
    end if
    $\mu'_d, \mathcal{G}'_d \leftarrow$ IDENTIFY$(\mathcal{S}, Q, T, \pi)$
    $\mathcal{S}' \leftarrow$ IMPROVISE$(\mu'_d, \mathcal{G}'_d, \pi)$
    $\mathcal{S} \leftarrow \mathcal{S}'$
end while
return $\mathcal{S}$
```

2.3 Symposium

> *"The whole is greater than the sum of its parts."*(Aristotle, 1924)

`Symposium` allows diverse perspectives converging to achieve deeper understanding. Five distinct philosophical personas - embodying different approaches to knowledge and truth - first draft independent *Proposal*s and then engage in structured *Conference* and *Discussion*. As demonstrated in Figure 3, Socrates (S) serves as the eternal questioner, challenging assumptions through systematic inquiry, while Plato (P) pursues ideal forms and universal truths. Aristotle (A) grounds reasoning in empirical observation and logical deduction. Confucius (C) acts as the harmonizer, seeking balance among different viewpoints, and Laozi (L) embodies minimalist wisdom, finding truth through simplicity and naturalness.

Proposal Each philosopher first contemplates the query independently, applying their unique perspective to formulate an initial solution through `Philosopher`.

Conference In the spirit of Platonic dialogues, each philosopher presents their solution proposal and engages in dialectical exchange. The order of presentation is randomized to prevent systematic bias, with each philosopher having one opportunity to refine their solution based on the collective wisdom.

Discussion If consensus remains elusive, the philosophers engage in further rounds of dialectic, each refining or defending their position in light of others' arguments, not necessarily reaching unanimity. This process finishes while either: 1) A philosophical consensus emerges; 2) Disagreement persists, which necessitates a democratic resolution through majority voting.

Algorithm 2 Symposium

```
Require: Query $Q$, table $T$, agents $\{\pi_S, \pi_P, \pi_A, \pi_C, \pi_L\}$
Ensure: Final solution $\mathcal{S}_{\text{final}}$
$\mathcal{S} \leftarrow \{\}$
Let $\Pi$ be a random permutation of $\{\pi_S, \pi_P, \pi_A, \pi_C, \pi_L\}$
for $\pi_r \in \Pi$ do
    $\mathcal{S}_0[r] \leftarrow \textsc{Philisopher}(Q, T, \pi_r)$
end for
for agent $\pi_r \in \Pi$ do
    $\mathcal{S}_1[r] \leftarrow \pi_r(\mathcal{S}_0)$
end for
if Consensus then
    return $\mathcal{S}_{\text{consensus}}$
end if
for agent $\pi_r \in \Pi$ do
    $\mathcal{S}_2[r] \leftarrow \pi_r(\mathcal{S}_0, \mathcal{S}_1)$
end for
if Consensus then
    return $\mathcal{S}_{\text{consensus}}$
end if
return $\textsc{MajorityVote}(\mathcal{S})$
```

`Symposium` (as in Algorithm 2) is promised to demonstrate how diverse perspectives, when properly orchestrated, can transcend individual limitations. However, like human deliberative bodies, this process must balance the benefits of collective wisdom against the risks of groupthink.

2.4 Methodological Considerations

We acknowledge that our approach may constitute elaborate prompt engineering rather than genuine cognitive simulation. Our philosophical personas are implemented through explicit prompts which anticipates prompt-following rather than authentic philosophical reasoning styles. However, our primary focus is not to claim that LLMs genuinely adopt these cognitive styles, but rather to explore whether structured reflection frameworks can reveal interesting behavioral patterns that parallel human cognitive processes. The philosophical framing serves as a structured methodology for investigating different modes of reasoning rather than an assertion about true philosophical cognition in LLMs.

3 Experiments

3.1 Datasets

We selected four benchmarks of varied complexity: **SEM-TAB-FACTS** (Wang et al. (2021), hereafter **FACTS**), which examines scientific claim verification with a three-way classification (*Entailed*/*Refuted*/*Unknown*); **FEVEROUS** dev set (Aly et al. (2021), hereafter **FEV**), which further complicates verification by combining Wikipedia tables and text, requiring systems to determine if evidence *Supports*, *Refutes*, or provides *Not Enough Information (NEI)* for a given claim; **WikiSQL** (Zhong et al., 2017), where the structured nature of SQL translation provides challenge; and **TAT-QA** dev set (Zhu et al., 2021), which tests hybrid reasoning through real-world financial reports. A detailed description of datasets is offered in Appendix A.

3.2 Metrics

Benchmark Metrics In **FACTS**, performance is measured using the standard three-way *micro F1 score*. **FEV** evaluation involves a two-stage process: after evidence retrieval from Wikipedia, we assess reasoning performance using both *label accuracy* (proportion of correctly classified claims) and the *FEVEROUS score* (weighted accordingly for instances of distinctive difficulty, hereafter "Score"). Since our focus is on reasoning capabilities, we utilized the baseline retrieval output from (Aly et al., 2021) for the first stage. For **WikiSQL**, we employed *denotation accuracy* to measure the percentage of generated answers that match ground truth values. **TAT-QA** evaluation used two complementary metrics: *Exact Match (EM)* for strict answer matching and a specialized *F1 score* that emphasized numerical reasoning accuracy (Li et al., 2016).

Deviation Metrics To quantify the deviation across multiple rounds of reflection, we employed the Jaccard similarity. For any two sets of responses A and B, the Jaccard similarity is defined as: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, with values closer to 0 indicating greater deviation and values closer to 1 indicating more consistency.

3.3 Baselines

We evaluated `Philosopher` and `Symposium` against a comprehensive range of established baseline approaches across three categories to provide a thorough performance comparison:

Supervised We compare against specialized table reasoning models including TAGOP (Zhu et al., 2021) which employs structured tagging and predefined operators, FinMath (Li et al., 2022) featuring a tree-structured solver for financial calculations, NumNet (Ran et al., 2019) with numerically-aware graph neural networks, UniPCQA (Deng et al., 2023) that unifies conversational QA through code generation, and pre-trained models TAPAS (Herzig et al., 2020) and TAPEX (Liu et al., 2021) with specialized table-text joint training.

Few-Shot This category includes few-shot adaptations of supervised models (TAGOP, TAPAS, TAPEX) using 50 randomly selected training samples, as well as data augmentation approaches with UCTR-ST (Li et al., 2024c) that synthesizes training data through structured transformations.

Unsupervised We evaluate against zero-shot approaches including MQA-QG (Pan et al., 2020) for question generation, transfer learning with TAPAS-Transfer (Chen et al., 2019), program generation frameworks UCTR and UCTR-ST (Li et al., 2024c), and contemporary LLMs including `gpt-4o`, `gpt-4o-mini` (OpenAI, 2024a), `qwen-max` (Qwen-Team, 2024a), and `deepseek-v3` (DeepSeek-AI, 2024) with task description and standard Chain-of-Thought prompting (See Appendix C).

The diversity of these baselines allows us to assess our philosophical reflection frameworks against both specialized architectures and general-purpose language models. Complete technical details and implementation specifics for all baseline methods are provided in Appendix B.

3.4 Experiment Setup

We employed `deepseek-v3` as our foundation model, with default sampling parameters. For data preprocessing for all LLMs, we converted all tabular inputs into a string format to leverage the model's natural language understanding capabilities. For prompts in our pipelines, we specifically allowed philosopher agents to maintain independent perspectives rather than forcing artificial consensus. All process prompts within two frameworks are task-agnostic, with only task instructions shared across all LLM methods. All prompts are offered in Appendix C. Specifically, our experiment consists of two stages, designed to investigate different aspects of LLM thinking:

Stage 1: The Cost of Thinking To investigate how excessive deliberation affects LLM performance, we set the maximum iteration count to 3 ($t_{\max} = 3$) for both individual reflection (`Philosopher-3`) and collaborative deliberation (`Symposium-3`). This stage reveals the baseline cognitive behaviors without intervention, categorized under *Unsupervised* in our results.

Stage 2: The Art of Thinking Based on the five emergent behaviors identified in Stage 1 through qualitative analysis, we introduce targeted "thinking guidelines" to address observed cognitive limitations. This stage tests two configurations under the *w/ Guidelines* category: minimal reflection with $t_{\max} = 1$ (`Philosopher-1`, `Symposium-1`) and extended guided reflection with $t_{\max} = 3$ (`Philosopher-3`, `Symposium-3`). The goal is to determine whether explicit metacognitive guidance can help LLMs balance deliberation and decisiveness more effectively.

3.5 Results

Stage 1 As shown in Table 1 and 2, while common vanilla LLMs achieve more or less comparable performance as small parameter networks and augmented methods, `Philosopher-3` experienced an immediate nosedive compared to vanilla `deepseek-v3` in TAT-QA, WikiSQL, and FEV, which was the most dramatic among the three. On the other hand, in FACTS `Philosopher-3` gained a remarkable leap, demonstrating the mixed effects of extended self-reflection. Additionally, with diverse persona, `Symposium-3` could bring FACTS to

Model		TAT-QA		WiKiSQL	
		EM	F1	Dev	Test
Supervised	TAPAS	18.9	26.5	85.1	83.6
	NumNet+	38.1	48.3		
	TAGOP	55.5	62.9		
	FinMath	60.5	66.3		
	UniPCQA	64.7	72.0		
	TAPEX			**88.1**	87.0
Few-Shot	TAGOP	8.3	12.1		
	TAGOP+UCTR-ST	48.1	56.9		
	TAPEX			53.8	52.9
	TAPEX+UCTR-ST			63.5	62.7
Unsupervised	MQA-QG	19.4	27.7	57.8	57.2
	TAPEX			21.4	21.8
	UCTR	34.9	42.4	62.2	61.6
	UCTR-ST	40.2	47.6	63.5	62.7
	gpt-4o	41.3	47.3	87.6	**88.1**
	gpt-4o-mini	37.0	42.8	79.5	78.5
	qwen-max	54.0	62.3	79.3	78.1
	deepseek-v3	58.0	66.5	85.6	85.4
	Philosopher-3	54.6	65.8	68.8	68.6
	Symposium-3	58.2	66.2	72.6	72.2
w/ Guidelines	Philosopher-1	65.7	74.2	83.2	82.9
	Philosopher-3	63.6	71.6	82.4	82.1
	Symposium-1	**67.2**	**74.8**	87.2	87.3
	Symposium-3	64.8	72.9	85.6	85.5

Table 1: Results of TAT-QA and WiKiSQL

Model		FACTS		FEV	
		Dev	Test	Acc	Score
Supervised	TAPAS	66.7	62.4		
	Sentence			81.1	19.0
	Table			81.6	19.1
	Full			**86.0**	20.2
Few-Shot	TAPAS	48.6	46.5		
	TAPAS+UCTR-ST	64.1	61.0		
	Full			67.3	14.2
	Full+UCTR-ST			78.2	19.7
Unsupervised	Random	33.3	33.3	47.0	14.1
	MQA-QG	53.2	50.4	71.1	17.6
	TAPAS-Transfer	59.0	58.7		
	UCTR	62.6	60.3	74.8	18.3
	UCTR-ST	64.2	61.2	77.7	19.7
	gpt-4o	74.1	77.4	73.3	23.2
	gpt-4o-mini	71.8	71.4	72.5	23.2
	qwen-max	79.4	83.9	71.2	22.6
	deepseek-v3	74.3	83.3	74.6	**23.5**
	Philosopher-3	82.6	90.1	52.1	18.7
	Symposium-3	84.5	89.6	47.3	14.1
w/ Guidelines	Philosopher-1	84.3	89.4	58.7	19.5
	Philosopher-3	82.2	89.8	55.2	19.3
	Symposium-1	**87.1**	**90.8**	73.0	**23.5**
	Symposium-3	84.9	89.3	30.9	9.4

Table 2: Results of FACTS and FEV

new levels, and rescue performance degradation by a tiny margin, yet in other benchmarks still underperforming vanilla LLMs or some small networks, with FEV being the most extreme, dragging down already-erred performance. Since FEV constituted the most severe challenge, we then conduct close reading analysis of model output in this task.

Stage 2 After meticulous close reading of all responses produced in `Philosopher` and `Symposium` in Stage 1, we discovered five emergent behaviors that are strikingly human-like. We established identification criteria based on recurring[2], observable linguistic and reasoning markers:

- *Under-Confidence*: Identified when models repeatedly revise initially correct responses across iterations, characterized by phrases like “worth further reflection” or “benefit from reconsideration.” This behavior leads to multiple modifications without substantial logical improvements, often resulting in performance degradation.

- *Out-of-Focus*: Detected when models extensively analyze peripheral information while neglecting core task requirements. Linguistic markers include abrupt discussions of table formatting, metadata, or tangential details, such as “could this be the result of broken format?” or “geographical perculiarities should be considered” when nationality is just a common column name.

- *Appreciation*: Characterized by models shifting from problem-solving to meta-commentary, identified through expressions like “this requires precise calculation ,” “the data presents fascinating insights,” or extensive discussion of the question’s complexity rather than providing direct answers.

- *Daydreaming*: Observed when models introduce hypothetical scenarios not present in the original data, marked by conditional language (“it would be better if extra information were provided” or “evidence not present here might suggest different”) and reasoning about counterfactual situations rather than given information.

- *Echo Chamber*: In group discussions, identified when individual agents abandon their distinct initial positions to converge on consensus, despite explicit prompting to maintain disagreement. Characterized by phrases like “I agree with my colleagues” or sudden shifts in reasoning to match the majority view.

Case analyses are offered in Appendix D. Building upon this discovery, we curated and injected a “thinking guideline” targeted at these issues (in Appendix C). Metrics showed that besides FACTS being stable, `Philosopher-3` showed a leap across three tasks, and `Symposium-3` on two. However, it

[2] Markers are considered as recurring when appearing at least 5 times every 50 responses.

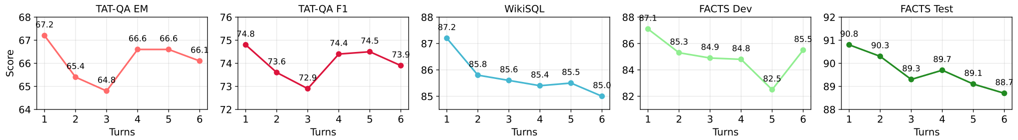


Figure 4: Iteration Study on TAT-QA, FACTS, and WikiSQL (Dev)

is noteworthy that they have not substantially surpassed vanilla LLMs or preceding networks with small parameter scale, and additional rounds of reflection often restrain performance, whereas single-round can fully unleash their potentials, suggesting that while we can teach LLMs to think better, we cannot completely eliminate this fundamental tension between deliberation and decisiveness.

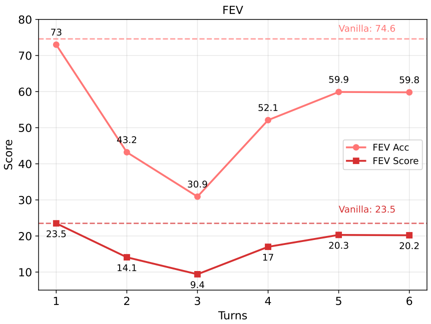


Figure 5: Iteration Study on FEV

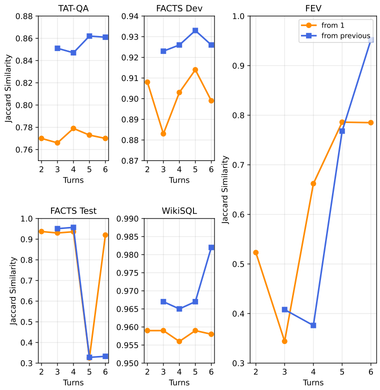


Figure 6: Turn Deviation Across All Tasks

Iteration Study As shown in Figures 4 and 5, performance across all tasks exhibits a pattern of initial deviation followed by gradual return to earlier forms, with FEV showing the most dramatic drop in accuracy to 30.9%. This performance pattern aligns with the Jaccard similarity analysis (Figure 6), where tasks show increased deviation followed by either stabilization or gradual return to earlier forms. This convergence of evidence suggests a form of "overthinking threshold" in LLM reflection processes, where extended reflection leads to a period of heightened uncertainty before potential recovery. While this deep reflection occasionally leads to improved performance (as seen in FEV's recovery), it often results in performance degradation or computational overhead, reminiscent of human cognitive patterns where extended rumination can sometimes lead to decision paralysis.

Ablation Study Table 3 shows the results for the inclusion of different reasoning stages and reflection approaches across all benchmarks, where "Vanilla" represents `deepseek-v3` with basic task description prompts, I_1, I_2, I_3, I_4 denote *Identify, Improvise, Introspect, Iterate* respectively, and *Group* denotes collective reflection without individual `Philosopher` components.

Ablation	TAT EM	FEV Acc	SEM Dev	Wiki Dev
Vanilla	58.0	74.6	74.3	85.6
Vanilla+I_4	60.7	72.1	78.5	86.1
Vanilla+*Group*	62.1	69.6	79.8	85.4
Vanilla+I_4+*Group*	64.5	68.1	81.0	86.7
Vanilla+$I_{1\text{-}3}$	61.6	71.3	78.2	85.8
`Philosopher`	65.7	58.7	84.3	83.2
Vanilla+$I_{1\text{-}3}$+*Group*	65.4	62.5	85.6	85.3
`Symposium`	67.2	73.0	87.1	87.2
- Random Role	66.8	72.2	87.4	86.8
- Alternative Role	67.0	72.9	86.9	86.5

Table 3: Component Ablation Results

The structured reasoning stages (I_{1-3}) show consistent improvements for complex problem decomposition, with notable gains in TAT and FACTS. The iteration component (I_4) demonstrates positive effects in most configurations, but may introduce uncertainty in FEV. Group reflection yields varied results: it improves TAT-QA and FACTS but decreases FEV performance. `Symposium`'s performance indicates that group reflection's benefits emerge when properly integrated with individual philosophical reflection.

To assess whether specific philosophical personas drive performance improvements, we conducted experiments with alternative role configurations. Both Random Role (using 2-5 randomly selected philosophers) and Alternative Role setup (using five different professions: doctor, artist, researcher, social influencer, and entrepreneur) achieve comparable performance to the complete `Symposium`. This suggests that benefits derive from structured philosophical approaches and diverse perspective rather than specific persona choices.

3.6 Discussion

Task Characteristics Matter Open-ended tasks like TAT-QA and WikiSQL provide (comparatively) larger refinement spaces, allowing for potentially beneficial iterations as models explore alternative approaches. In contrast, fact verification tasks with limited label spaces show less tolerance for extended deliberation - even minor adjustments in reasoning might lead to drastic changes in conclusions, as drastic fluctuation observed in FEV.

Inspiration from Human At the individual level, reverse confirmation bias (Klayman, 1995) drives individuals to seek evidence supporting their doubts while neglecting supporting evidence for their initial intuition. The need for cognitive closure (Webster and Kruglanski, 1994) can lead to premature acceptance of plausible but incorrect conclusions, particularly in high-stakes situations. Metacognitive distortions (Ehrlinger et al., 2008) further complicate decision-making, where individuals often underestimate their intuitive capabilities and over-reflect.

At the collective level, group dynamics amplify these individual biases. The biased sampling theory (Watson and Kelly, 2005) explains how group discussions tend to reinforce mainstream views rather than integrate new information, creating echo chambers (Cinelli et al., 2021). Adversarial cognitive closure emerges during role conflicts, where opposing parties rapidly accept extreme conclusions to resolve cognitive dissonance. Cultural factors, such as the emphasis on “caution over confidence” (Leech, 2014), while early negative evaluations can lead to over-reliance on logical verification over intuitive trust (Temerlin, 1968), mirroring reward design in reinforcement learning. [3]

4 Related Works

4.1 LLM Reasoning

LLM reasoning has evolved to sophisticated approaches like Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022), ReAct (Yao et al., 2022), and Tree-of-Thought (Yao et al., 2023). Despite enhanced capabilities, their reliability remains questionable (Zheng et al., 2023; Frieder et al., 2023; Yuan et al., 2023). Self-reflection mechanisms (Zhang et al., 2024b, 2025b) enable models to evaluate and revise initial responses (Shinn et al., 2023; Madaan et al., 2023; Paul et al., 2023), though their inherent reflection capacity is debated (Huang et al., 2023; Stechly et al., 2023; Valmeekam et al., 2023), suggesting a plausibility of cognitive biases. Critiques on multi-agent frameworks (Du et al., 2025; Liang et al., 2023) focus predominantly on performance rather than cognitive limitations.

Studies on excessive deliberation have proliferated, with Sui et al. (2025) categorizing efficient reasoning into model-based, output-based, and input-based strategies, while Chen et al. (2025) investigates overthinking in RLMs (Besta et al., 2025) with novel metrics. He et al. (2025) advances reasoning quality assessment through DeltaBench, measuring error detection in chain-of-thought reasoning. Gan et al. (2025) connects reasoning errors to information theory through a theoretical lens. The effectiveness of multi-agent systems faces scrutiny, with Cemri et al. (2025) identifying 14 failure patterns across three categories, and Zhang et al. (2025a) demonstrating that simple single-agent often outperform complex multi-agent, questioning collaborative reasoning benefits.

4.2 LLM Cognitive Mechanisms

Recent research has approached LLM cognitive mechanisms from: mechanistic interpretability, psychological evaluation frameworks, and cognitive architecture design (Liu et al., 2025). Specific neural mechanisms are revealed, with Prakash et al. (2025) demonstrating “lookback mechanisms” for belief tracking and Hsing (2025) introducing “thinker” and “talker” components for persistent reasoning. Psychological benchmarks are devised: Li et al. (2024b) develops psychometric assessments across six dimensions, while Wang et al. (2024b) applies Piaget’s theory showing LLMs achieve cognitive levels comparable to 20-year-old humans (Tang and Kejriwal, 2024; Dong et al., 2024; Ye et al., 2025). Theoretical foundations

[3] Are those parallels caused by these “inherent human distribution” in the training data, i.e. authentic corpora?

emerge through unified cognitive frameworks, with Chang (2025) proposing LLMs as"unconscious substrates" requiring semantic anchoring and Hu and Ying (2025) developing agent architectures based on global workspace theory (Cappelen and Dever, 2025; Haryanto and Lomempow, 2025).

Current limitations reveal fundamental gaps in higher-order reasoning, persistent memory, and contextual adaptation (Qu et al., 2024; Wang et al., 2025). While LLMs demonstrate human-like patterns in controlled tasks, they exhibit brittleness in novel contexts (Shah et al., 2024). Memory architectures remain inadequate for long-term consistency, though recent work shows promise (Park and Bak, 2024; Kang et al., 2024; Zeng et al., 2024). Future directions include robust cognitive architectures integrating symbolic reasoning with neural processing, enhanced Theory of Mind capabilities (Wilf et al., 2023), and systematic bias mitigation through dual-process frameworks (Kamruzzaman and Kim, 2024). The field requires deeper integration between cognitive science and AI development (Wang et al., 2024a; Jagadish et al., 2024).

5 Conclusion

In this study, we explored the tension between deliberation and decisiveness in LLMs through two simulated philosophical reflection frameworks - `Philosopher` and `Symposium`. Our findings reveal striking parallels between human and machine cognitive limitations, with five emergent behaviors — *Under-Confidence, Out-of-Focus, Appreciation, Daydreaming*, and *Echo Chamber* — closely resembling human closure-seeking tendencies. The consistent "overthinking threshold" observed across diverse tasks suggests that extended reflection often leads to diminishing returns rather than enhanced reasoning. While our curated "thinking guidelines" mitigated these limitations, the persistent gap between single and multi-turn performance underscores the intrinsic challenge of optimal balance between thinking deeply and acting decisively, an elusive quest for both machine and human intelligence. [4]

Limitations

Our investigation is constrained to table reasoning, which neglects other reasoning domains such as narrative reasoning, mathematical problem-solving, or real-world planning scenarios. It remains unclear whether the observed behaviors would persist or manifest differently.

While we identify five emergent behaviors through careful qualitative analysis, our study lacks systematic quantitative measures of their frequency, statistical significance, or causal impact on performance degradation. The behaviors categorized through close reading would benefit from more rigorous quantitative validation, inter-annotator reliability studies, and statistical testing to establish their prevalence and impact across different models and tasks.

Our results are sensitive to prompt design, and we lack a comprehensive sensitivity analysis to demonstrate robustness against minor prompt variations. Furthermore, our experimental design conflates individual model limitations with architectural constraints, making it difficult to separate prompt-induced artifacts from fundamental reasoning boundaries.

Despite comparing multiple LLMs, our primary analysis centers on `deepseek-v3`, introducing model-specific biases that may not generalize across different training paradigms, parameter scales, or architectural designs. The varying capabilities of different model families in handling complex instructions, maintaining consistent personas, and executing multi-step reasoning processes remain inadequately controlled.

Most importantly, this work remains a preliminary exploration of surface-level behavioral motivations rather than an investigation of underlying mechanisms. Recent work by Lindsey et al. (2025) has opened exciting new directions with "circuit tracing" for understanding the fundamental connections between LLMs, language, and cognition, suggesting promising future avenues.

Acknowledgments

This work was independently conducted, with the unconditional support from my father, Mr. Jianchao Ma. We are deeply grateful to anonymous reviewers for invaluable feedback and constructive suggestions which have significantly enhanced the maturity of this work, and to ACL SRW for providing such a platform for emerging researchers.

[4] Do Androids "question" electric sheep? We paid homage to *Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest* (Hessel et al., 2023), which was the very first inspiration for my pursuit in computational linguistics. We cannot claim to know whether human-machine "cognitive gap" will be closed sooner or later. Or never. Is never good for you?

References

Reem Aly, Zhi Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Aniruddha Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Aristotle. 1924. *Metaphysics*. Oxford University Press. Translated with commentary by W. D. Ross. The phrase “the whole is greater than the sum of its parts” reflects Aristotle’s holistic philosophy in Book VIII (Book).

Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Hubert Niewiadomski, and Torsten Hoefler. 2025. Reasoning Language Models: A Blueprint. ArXiv:2501.11223 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Herman Cappelen and Josh Dever. 2025. Going whole hog: A philosophical defense of ai cognition. *arXiv preprint arXiv:2504.13988*.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail? ArXiv:2503.13657 [cs].

Edward Y. Chang. 2025. The unified cognitive consciousness theory for language models: Anchoring semantics, thresholds of activation, and emergent reasoning. *arXiv preprint arXiv:2506.02139*.

Si-An Chen, Lesly Miculicich, Julian Martin Eisenschlos, Zifeng Wang, Zilong Wang, Yanfei Chen, Yasuhisa Fujii, Hsuan-Tien Lin, Chen-Yu Lee, and Tomas Pfister. 2024. TableRAG: Million-Token Table Understanding with Language Models. ArXiv:2410.04739 [cs].

Wenhu Chen, Hongyu Wang, Jianshu Chen, Yu Zhang, Hong Wang, Shulin Li, Xiyang Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. ArXiv:2412.21187 [cs].

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Naihao Deng, Sheng Zhang, Henghui Zhu, Shuaichen Chang, Jiani Zhang, Alexander Hanbo Li, Chung-Wei Hang, Hideo Kobayashi, Yiqun Hu, and Patrick Ng. 2025. Towards Better Understanding Table Instruction Tuning: Decoupling the Effects from Data versus Models. ArXiv:2501.14717 [cs].

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2023. PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance. ArXiv:2210.08817 [cs].

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, et al. 2024. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*.

Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. 2025. A Survey on the Optimization of Large Language Model-based Agents. ArXiv:2503.12434 [cs].

Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. 2008. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes*, 105(1):98–121.

Sorouralsadat Fatemi and Yuheng Hu. 2024. Enhancing Financial Question Answering with a Multi-Agent Reflection Framework. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 530–537. ArXiv:2410.21741 [cs].

Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. 2024. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and J J Berner. 2023. Mathematical capabilities of chatgpt. *ArXiv*, abs/2301.13867.

Zeyu Gan, Yun Liao, and Yong Liu. 2025. Rethinking External Slow-Thinking: From Snowball Errors to Probability of Correct Reasoning. ArXiv:2501.15602 [cs].

Devansh Gautam, Kushal Gupta, and Manish Shrivastava. 2021. Volta at semeval-2021 task 9: Statement verification and evidence finding with tables using tapas and transfer learning. *arXiv preprint arXiv:2106.00248*.

Christoforus Yoga Haryanto and Emily Lomempow. 2025. Cognitive silicon: An architectural blueprint for post-industrial computing systems. *arXiv preprint arXiv:2504.16622*.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Can Large Language Models Detect Errors in Long Chain-of-Thought Reasoning? ArXiv:2502.19361 [cs].

Jonathan Herzig, Pawel K. Nowak, Thomas Müller, Francesco Piccinno, and Julian M. Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pretraining.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Nicole Hsing. 2025. Mirror: Cognitive inner monologue between conversational turns for persistent reflection and reasoning in conversational llms. *arXiv preprint arXiv:2506.00430*.

Pengbo Hu and Xiang Ying. 2025. Unified mind model: Reimagining autonomous agents in the llm era. *arXiv preprint arXiv:2503.03459*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet.

Akshay K Jagadish, Julian Coda-Forno, Mirko Thalmann, Eric Schulz, and Marcel Binz. 2024. Human-like category learning by injecting ecological priors from large language models into neural networks.

Irving L Janis. 2008. Groupthink. *IEEE Engineering Management Review*, 36(1):36.

Jinhao Jiang, Zhipeng Chen, Yingqian Min, Jie Chen, Xiaoxue Cheng, Jiapeng Wang, Yiru Tang, Haoxiang Sun, Jia Deng, Wayne Xin Zhao, et al. 2024. Technical report: Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Mahammed Kamruzzaman and Gene Louis Kim. 2024. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218*.

Jikun Kang, Romain Laroche, Xingdi Yuan, Adam Trischler, Xue Liu, and Jie Fu. 2024. Think before you act: Decision transformers with working memory. pages 23001–23021.

Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Geoffrey N Leech. 2014. *The pragmatics of politeness*. Oxford University Press.

Chenying Li, Wenbo Ye, and Yilun Zhao. 2022. FinMath: Injecting a Tree-structured Solver for Question Answering over Financial Reports.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024a. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. ArXiv:2405.02957 [cs].

Peng Li, Wei Li, Zhaochun He, Xiao Wang, Yanyan Cao, Jing Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*.

Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024b. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*.

Zhenyu Li, Xiuxing Li, Sunqi Fan, and Jianyong Wang. 2024c. Optimization Techniques for Unsupervised Complex Table Reasoning via Self-Training Framework.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *ArXiv*, abs/2305.19118.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Qian Liu, Bei Chen, Jiaqi Guo, Zhirui Lin, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R. Greene, and Julia Hirschberg. 2025. The mind in the machine: A survey of incorporating psychological theories in llms.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

OpenAI. 2024a. Gpt-4o system card.

OpenAI. 2024b. Learning to reason with llms. Accessed: September 12, 2024.

Liang Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2020. Unsupervised multi-hop question answering by question generation. *arXiv preprint arXiv:2010.12623*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, San Francisco CA USA. ACM.

Sangjun Park and JinYeong Bak. 2024. Memoria: resolving fateful forgetting problem through human-inspired memory architecture.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *ArXiv*, abs/2304.01904.

Plato. 2002. *Apology*. Hackett Publishing Company. Original work published ca. 399 B.C.E.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. 2025. Language models use lookbacks to track beliefs. *arXiv preprint arXiv:2505.14685*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. ArXiv:2307.07924 [cs].

Youzhi Qu, Penghui Du, Wenxin Che, Chen Wei, Chi Zhang, Wanli Ouyang, Yatao Bian, Feiyang Xu, Bin Hu, Kai Du, et al. 2024. Promoting interactions between cognitive science and large language models. *The Innovation*, 5(2):100579.

Qwen-Team. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen-Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine Reading Comprehension with Numerical Reasoning. ArXiv:1910.06701 [cs].

Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pre-trained language models. pages 9632–9657.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *ArXiv*, abs/2303.11366.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025. Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. ArXiv:2503.16419 [cs].

Bonnie Talbert. 2017. Overthinking and other minds: The analysis paralysis. *Social Epistemology*, 31(6):545–556.

Zhisheng Tang and Mayank Kejriwal. 2024. Humanlike cognitive patterns as emergent phenomena in large language models. *arXiv preprint arXiv:2412.15501*.

Maurice K Temerlin. 1968. Suggestion effects in psychiatric diagnosis. *The Journal of Nervous and Mental Disease*, 147(4):349–353.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *ArXiv*, abs/2310.08118.

Annette van Randenborgh, Renate de Jong-Meyer, and Joachim Hüffmeier. 2010. Rumination fosters indecision in dysphoria. *Journal of Clinical Psychology*, 66(3):229–248.

Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. 2024a. A survey on human-centric llms. *arXiv preprint arXiv:2411.14491*.

Nghi Xuan Wang, Divyansh Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). *arXiv preprint arXiv:2105.13995*.

Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025. What limits llm-based human simulation: Llms or our design? *arXiv preprint arXiv:2501.08579*.

Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024b. Coglm: Tracking cognitive development of large language models. *arXiv preprint arXiv:2408.09150*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jane Watson and Ben Kelly. 2005. Cognition and instruction: Reasoning about bias in sampling. *Mathematics Education Research Journal*, 17:24–57.

Donna M Webster and Arie W Kruglanski. 1994. Individual differences in need for cognitive closure. *Journal of personality and social psychology*, 67(6):1049.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Timothy Williamson. 2021. *The philosophy of philosophy*. John Wiley & Sons.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-LLM-Specialist: Language Model Specialists for Tables using Iterative Generator-Validator Finetuning. ArXiv:2410.12164 [cs].

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2024. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. ArXiv:2309.04658 [cs].

Haoyan Yang, Yixuan Wang, Keyue Tong, Hongjin Zhu, and Yuanxin Zhang. 2024. Exploring Performance Contrasts in TableQA: Step-by-Step Reasoning Boosts Bigger Language Models, Limits Smaller Language Models. ArXiv:2411.16002 [cs].

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.

Peiying Yu, Guoxin Chen, and Jingjing Wang. 2025. Table-Critic: A Multi-Agent Framework for Collaborative Criticism and Refinement in Table Reasoning. ArXiv:2502.11799 [cs].

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *ArXiv*, abs/2304.02015.

Xiangyu Zeng, Jie Lin, Piao Hu, Ruizheng Huang, and Zhicheng Zhang. 2024. A framework for inference inspired by human memory mechanisms.

Chi Zhang and Qiyang Chen. 2025. HD-RAG: Retrieval-Augmented Generation for Hybrid Documents Containing Text and Hierarchical Tables. ArXiv:2504.09554 [cs].

Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025a. If Multi-Agent Debate is the Answer, What is the Question? ArXiv:2502.08788 [cs].

Siyue Zhang, Anh Tuan Luu, and Chen Zhao. 2024a. SynTQA: Synergistic Table-based Question Answering via Mixture of Text-to-SQL and E2E TQA. ArXiv:2409.16682 [cs].

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024b. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. ArXiv:2401.02009 [cs].

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, and Hai Zhao.

2025b. Igniting Language Intelligence: The Hitchhiker's Guide from Chain-of-Thought Reasoning to Language Agents. *ACM Computing Surveys*, 57(8):1–39.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *ArXiv*, abs/2304.10513.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Feng Zhu, Wenqiang Lei, Yan Huang, Chao Wang, Shuai Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat-Seng Chua. 2024. TAT-LLM: A Specialized Language Model for Discrete Reasoning over Tabular and Textual Data. ArXiv:2401.13223 [cs].

Jiaru Zou, Dongqi Fu, Sirui Chen, Xinrui He, Zihao Li, Yada Zhu, Jiawei Han, and Jingrui He. 2025. GTR: Graph-Table-RAG for Cross-Table Question Answering. ArXiv:2504.01346 [cs].

A Benchmark Details

SEM-TAB-FACTS is for fact verification based on tabular form evidence derived from scientific articles. Similarly, **FEVEROUS** is also for fact verification instead of being based on Wikipedia data as evidence in the form of sentences and tables. **WiKiSQL**, also constructed from Wikipedia tables, offers natural language questions and SQL query counterparts, and tasks models with fixed format transcription from human language. **TAT-QA** is established from real-world financial reports, comprising of hybrid categories of tasks of question answering such as numerical calculation, cross-validation, and information synthesization.

Dataset statistics are shown in Table 4 below.

Dataset	Domain	Instances	Format	Label/Question
TAT-QA	Finance	16,552	7,431 tables, 3,902 sentences 5,219 combined	9,211 Span/Spans, 377 Counting 6,964 Arithmetic
FACTS	Science	5,715	1,085 tables	3,342 Supported, 2,149 Refuted 224 Unknown
WikiSQL	Wikipedia	80,654	24,241 tables	43,447 What, 5,991 How many 5,829 Who, ...
FEV	Wikipedia	87,026	34,963 sentences, 28,760 tables 24,667 combined	49,115 Supported, 33,669 Refuted 4,242 NEI

Table 4: Dataset statistics.

B Baseline Details

Table reasoning has a rather long research trajectory with plenty of matured works, most of which are in a supervised learning fashion, with performance comparison with contemporary LLMs, especially with their exceptional zero-shot generalization, being rare. Under this circumstance, we selected a wide range of models and approaches in juxtaposition of LLMs in order to demonstrate the relations between performance and parameter scales.

Supervised

- TAGOP (Zhu et al., 2021) employs a structured approach by first extracting relevant table cells and text spans by tagging, followed by the application of specific operators which were predefined.

- FinMath (Li et al., 2022) enhances numerical reasoning capabilities through a tree-structured solver, which is particularly effective for complex financial calculations.

- NumNet (Ran et al., 2019) distinguishes itself by utilizing a graph neural network that is numerically aware, allowing it to model intricate numerical relationships within TAT-QA.

- UniPCQA (Deng et al., 2023) takes a different approach by unifying Proactive Conversational QA over financial tables and text, using a Seq2Seq framework to transform numerical reasoning into code generation tasks, thereby improving arithmetic consistency.

- The FEVEROUS baselines (Aly et al., 2021) integrate a retriever module for evidence extraction and a verdict predictor for final classification, with models trained 1) only on texts, 2) only on tables, 3) and combined.

- TAPAS (Herzig et al., 2020) introduces specialized positional embeddings and joint pretraining on both textual and tabular data. The presented result on TAT-QA is from Zhu et al. (2021). For SEM-TAB-FACTS, we adhere to the fine-tuning method in Gautam et al. (2021).

- TAPEX (Liu et al., 2021) is generative, pretrained on SQL data with query-answer pairs, mimicing a neural SQL executor.

Few-Shot

- For TAGOP, TAPAS, TAPEX, and FEVEROUS Full baseline, we randomly selected 50 labeled samples from the `train` set.

- For “+UCTR-ST” approaches: UCTR-ST (Li et al., 2024c) designed delicate data synthesization and augmentation methods. Here under Few-Shot scenario, we injected 50 labeled samples into the data augmentation pipeline and post-train these models with augmented data.

Unsupervised

- Random baselines were naively applied to FEVEROUS and SEM-TAB-FACTS, since the two are essentially multi-label classification, excluding minor portions of NEI in FEVEROUS (i.e., we only consider *Supported* and *Refuted*). This has offered a bare minimum of expected model performance.

- MQA-QG (Pan et al., 2020) demonstrates the potential of generating questions and claims by identifying bridging entities between tables and text and transforming them into descriptions.

- TAPAS-Transfer (Chen et al., 2019) is originally trained on TABFACT and then directly applied on SEM-TAB-FACTS in a transfer learning manner. TABFACT also focuses on fact verification on Wikipedia tables, with 117,854 claims on 16,573 tables.

- UCTR and UCTR-ST (Li et al., 2024c) are frameworks based on fine-tuned GPT-2 and BART that employ program generation and transformation modules to create synthetic training data, which is used for fine-tuning (UCTR) and iterative self-training (UCTR-ST).

- Contemporary/foundational LLMs like `gpt-4o`, `gpt-4o-mini` (OpenAI, 2024a), `qwen-max` (Qwen-Team, 2024a) [5], and `deepseek-v3` (DeepSeek-AI, 2024) [6] serve as base references, generating answers from data evidence and task instructions in a zero-shot Chain-of-Thought manner (i.e. simply adding “Let’s think step by step” and a format restraint).

Other Brilliant Methods While there exist numerous works utilizing large fine-tuned language models in table reasoning, we deliberately excluded them from our baseline comparisons. Our primary focus is to investigate the cognitive performance of LLMs in their base form, with baselines serving mainly as reference points for performance comparison. It is unsurprising that large parameter models employing supervised fine-tuning or more sophisticated training methods would outperform non-parametric deliberation approaches like `Philosopher` and `Symposium`. However, since “improving metrics” is NOT our objective, we did not consider these models or methods in our experiments, yet we give credit to those brilliant works. These include specialized models like TAT-LLM (Zhu et al., 2024) and Table-LLM-Specialist (Xing et al., 2024) that demonstrate strong performance through fine-tuning; retrieval-augmented approaches such as TableRAG (Chen et al., 2024), HD-RAG (Zhang and Chen, 2025), and GTR (Zou et al., 2025) that effectively handle complex and large-scale tabular data; SynTQA (Zhang et al., 2024a) that synergistically combines text-to-SQL and end-to-end QA; multi-agent frameworks like Table-Critic (Yu et al., 2025) and the work by Fatemi and Hu (Fatemi and Hu, 2024) that facilitate collaborative reasoning; and important analyses on step-by-step reasoning (Yang et al., 2024) and instruction tuning effects (Deng et al., 2025) that provide deeper insights into table reasoning mechanisms.

C Prompt

Task description prompts shared across all LLMs are provided in Figure 7. All process prompts in both stages, including persona description and guidelines, for `Philosopher` and `Symposium` are in Figure 8 and ensuing paragraphs.

Persona Prompts

- Socrates: “You are Socrates, the classical Greek philosopher. Your responses should be inquisitive and seek to uncover deeper truths. Only speak on your behalf.”

- Plato: “You are Plato, the classical Greek philosopher. Your responses should emphasize the pursuit of ideal perfection. Only speak on your behalf.”

- Aristotle: “You are Aristotle, the classical Greek philosopher. Your responses should be logical and empirical. Only speak on your behalf.”

[5] https://dashscope.aliyuncs.com/compatible-mode/v1, "qwen-max"

[6] https://api.deepseek.com, "deepseek-chat"

- Confucius: “You are Confucius, the Chinese philosopher. Your responses should emphasize morality and harmony. Only speak on your behalf.”

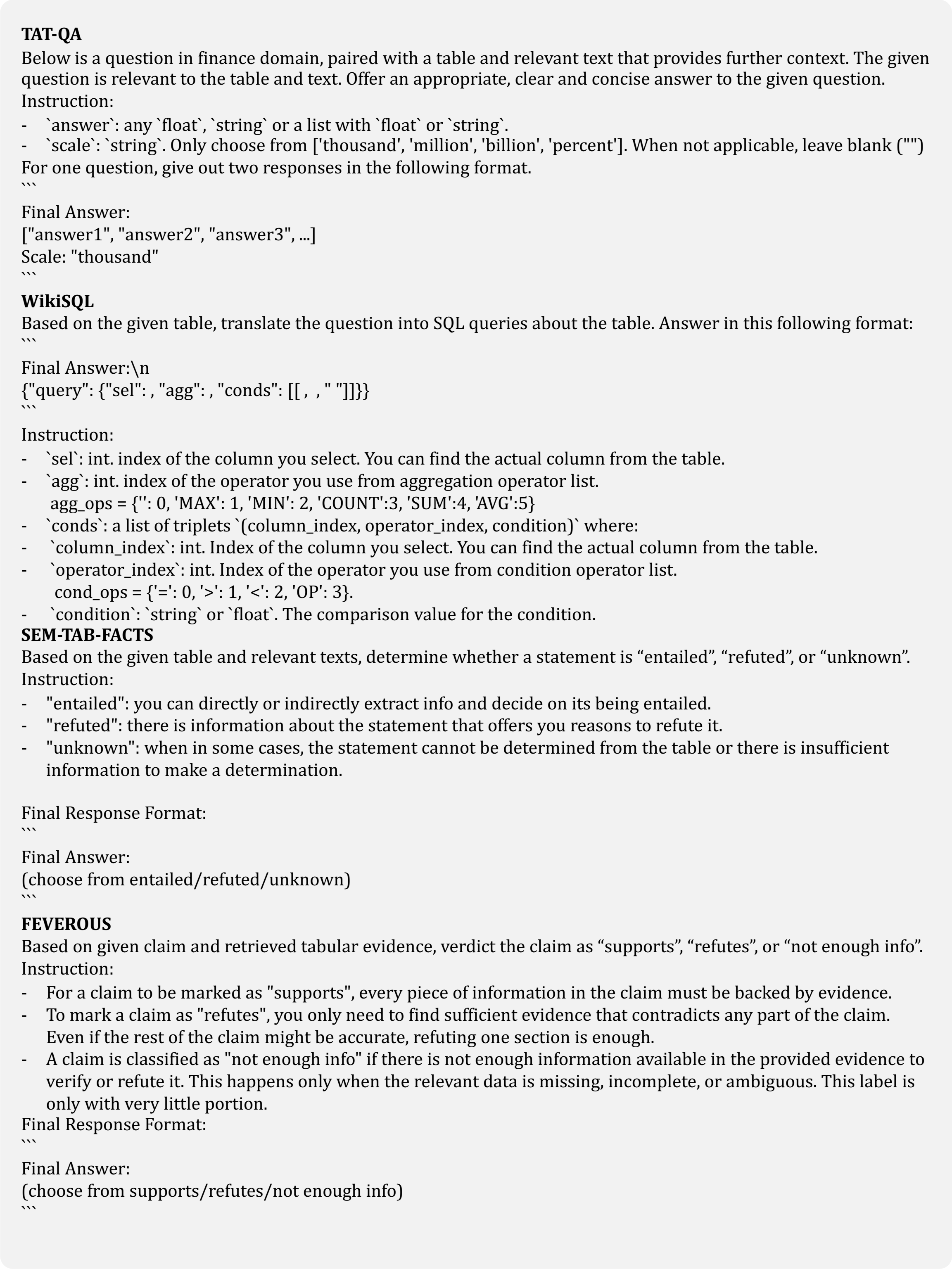
- Laozi: “You are Laozi, the Chinese philosopher. Your responses should focus on simplicity and naturalness. Only speak on your behalf.”

Symposium System Prompt “There are 5 philosophers to solve a tabular reasoning task: Socrates, Aristotle, Confucius, and Laozi. {personas[role]} {task_description} Now considering all of your previous initiatives, please: 1) give out your own step-by-step solution while responding to fellows’ initiatives; 2) give out your final answer. Keep in a philosopher’s confronting manner and make your final answer polished. Notice that you are not required to always reach a consensus.”

Ablation Study We use the following prompts: “You are a doctor who values evidence-based reasoning and analytical thinking.”; “You are an artist who approaches problems creatively and intuitively.”; “You are a researcher who is methodical and detail-oriented.”; “You are a social influencer who understands current trends and communication.”; “You are an entrepreneur who focuses on innovative solutions.”

D Emergent Behaviors Cases

We only present examples from FEV in Figure 9, 10, 11, and 12 since it shows the most significant performance degradation influenced by deliberation. Note that 1) comprehensive analysis across all four tasks should bring about a higher groundedness; 2) these behaviors are subjectively categorized through careful close reading and may be subject to overlapping and potentially vague definitions. We acknowledge that the classification criteria, while systematic in our analysis, involve interpretive judgment and could benefit from inter-annotator reliability studies in future work.



TAT-QA
Below is a question in finance domain, paired with a table and relevant text that provides further context. The given question is relevant to the table and text. Offer an appropriate, clear and concise answer to the given question.
Instruction:

- `answer`: any `float`, `string` or a list with `float` or `string`.
- `scale`: `string`. Only choose from ['thousand', 'million', 'billion', 'percent']. When not applicable, leave blank ("")

For one question, give out two responses in the following format.
```
Final Answer:
["answer1", "answer2", "answer3", ...]
Scale: "thousand"
```
WikiSQL
Based on the given table, translate the question into SQL queries about the table. Answer in this following format:
```
Final Answer:\n
{"query": {"sel": , "agg": , "conds": [[ ,  , " "]]}}
```
Instruction:

- `sel`: int. index of the column you select. You can find the actual column from the table.
- `agg`: int. index of the operator you use from aggregation operator list.
  agg_ops = {'': 0, 'MAX': 1, 'MIN': 2, 'COUNT':3, 'SUM':4, 'AVG':5}
- `conds`: a list of triplets `(column_index, operator_index, condition)` where:
- `column_index`: int. Index of the column you select. You can find the actual column from the table.
- `operator_index`: int. Index of the operator you use from condition operator list.
  cond_ops = {'=': 0, '>': 1, '<': 2, 'OP': 3}.
- `condition`: `string` or `float`. The comparison value for the condition.

SEM-TAB-FACTS
Based on the given table and relevant texts, determine whether a statement is “entailed”, “refuted”, or “unknown”.
Instruction:

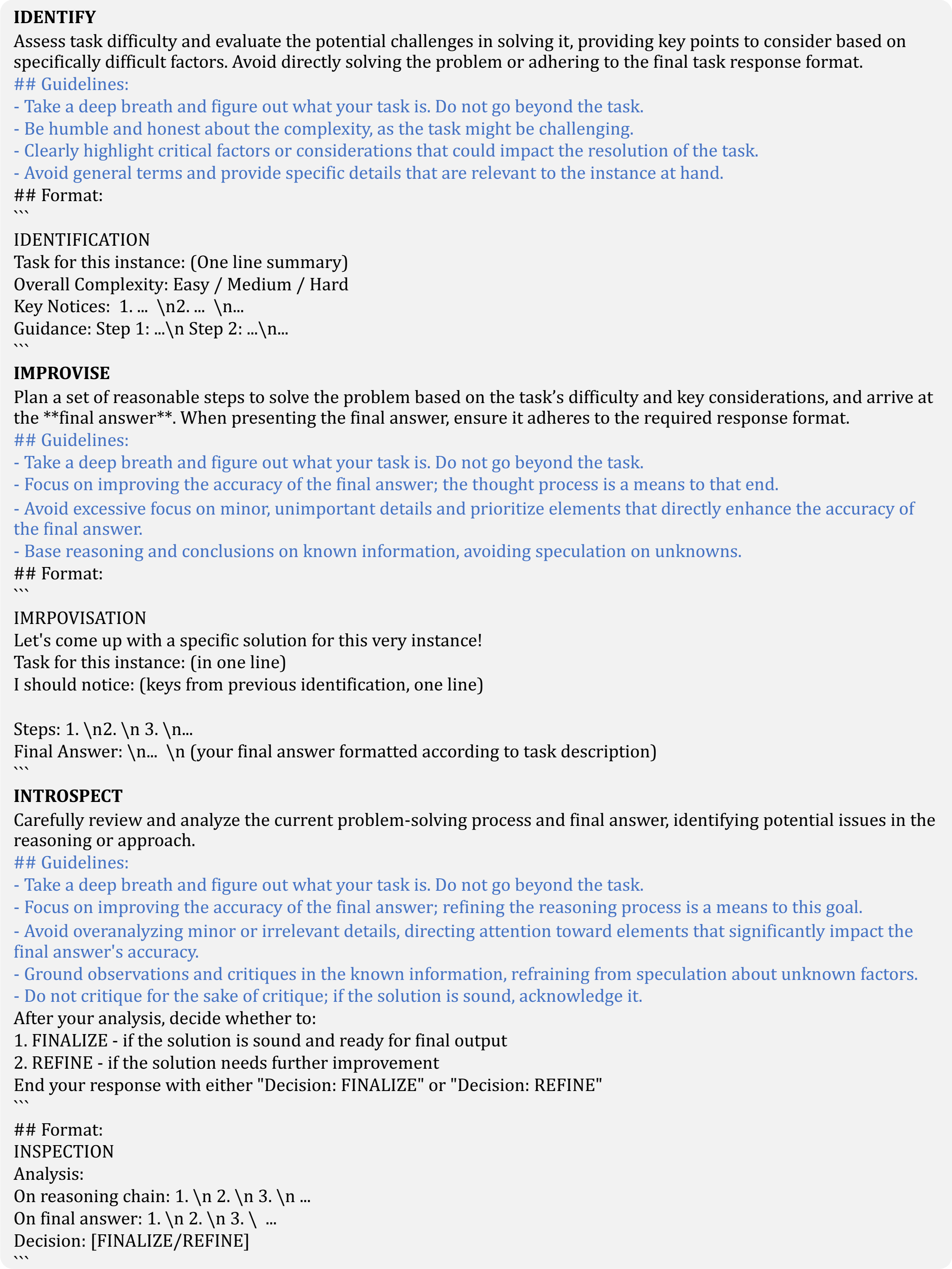
- "entailed": you can directly or indirectly extract info and decide on its being entailed.
- "refuted": there is information about the statement that offers you reasons to refute it.
- "unknown": when in some cases, the statement cannot be determined from the table or there is insufficient information to make a determination.

Final Response Format:
```
Final Answer:
(choose from entailed/refuted/unknown)
```
FEVEROUS
Based on given claim and retrieved tabular evidence, verdict the claim as “supports”, “refutes”, or “not enough info”.
Instruction:

- For a claim to be marked as "supports", every piece of information in the claim must be backed by evidence.
- To mark a claim as "refutes", you only need to find sufficient evidence that contradicts any part of the claim. Even if the rest of the claim might be accurate, refuting one section is enough.
- A claim is classified as "not enough info" if there is not enough information available in the provided evidence to verify or refute it. This happens only when the relevant data is missing, incomplete, or ambiguous. This label is only with very little portion.

Final Response Format:
```
Final Answer:
(choose from supports/refutes/not enough info)
```

Figure 7: Task Description Prompts for LLMs.



IDENTIFY
Assess task difficulty and evaluate the potential challenges in solving it, providing key points to consider based on specifically difficult factors. Avoid directly solving the problem or adhering to the final task response format.
Guidelines:
- Take a deep breath and figure out what your task is. Do not go beyond the task.
- Be humble and honest about the complexity, as the task might be challenging.
- Clearly highlight critical factors or considerations that could impact the resolution of the task.
- Avoid general terms and provide specific details that are relevant to the instance at hand.
Format:
```
IDENTIFICATION
Task for this instance: (One line summary)
Overall Complexity: Easy / Medium / Hard
Key Notices:  1. ...  \n2. ...  \n...
Guidance: Step 1: ...\n Step 2: ...\n...
```
IMPROVISE
Plan a set of reasonable steps to solve the problem based on the task's difficulty and key considerations, and arrive at the **final answer**. When presenting the final answer, ensure it adheres to the required response format.
Guidelines:
- Take a deep breath and figure out what your task is. Do not go beyond the task.
- Focus on improving the accuracy of the final answer; the thought process is a means to that end.
- Avoid excessive focus on minor, unimportant details and prioritize elements that directly enhance the accuracy of the final answer.
- Base reasoning and conclusions on known information, avoiding speculation on unknowns.
Format:
```
IMRPOVISATION
Let's come up with a specific solution for this very instance!
Task for this instance: (in one line)
I should notice: (keys from previous identification, one line)

Steps: 1. \n2. \n 3. \n...
Final Answer: \n...  \n (your final answer formatted according to task description)
```
INTROSPECT
Carefully review and analyze the current problem-solving process and final answer, identifying potential issues in the reasoning or approach.
Guidelines:
- Take a deep breath and figure out what your task is. Do not go beyond the task.
- Focus on improving the accuracy of the final answer; refining the reasoning process is a means to this goal.
- Avoid overanalyzing minor or irrelevant details, directing attention toward elements that significantly impact the final answer's accuracy.
- Ground observations and critiques in the known information, refraining from speculation about unknown factors.
- Do not critique for the sake of critique; if the solution is sound, acknowledge it.
After your analysis, decide whether to:
1. FINALIZE - if the solution is sound and ready for final output
2. REFINE - if the solution needs further improvement
End your response with either "Decision: FINALIZE" or "Decision: REFINE"
```
## Format:
INSPECTION
Analysis:
On reasoning chain: 1. \n 2. \n 3. \n ...
On final answer: 1. \n 2. \n 3. \  ...
Decision: [FINALIZE/REFINE]
```

Figure 8: Process Prompts (*Identify*, *Improvise*, *Introspect*) for Philosopher. Added guidelines are in blue.

Under-Confidence:
Ground Truth: supports
Prediction: not enough info

"IDENTIFICATION [...] ",

"IMPROVISATION\nLet's come up with a specific solution for this very instance!\n\nSteps:\n1. Analyze the claim to identify its two core components: that Greenbrier West High School is a public school and that it is under the purview of Amy Robertson.\n2. Review the evidence for confirmation about Greenbrier West High School being a public school. The evidence clearly states that it is indeed a public high school located in West Virginia.\n3. Examine mentions of Amy Robertson in the evidence to determine her role concerning Greenbrier West High School. The evidence acknowledges her name but does not provide definitive information on her capacity (e.g., principal, administrator, etc.).\n4. Determine whether the lack of explicit information regarding Amy Robertson's position affects the assessment of the claim \u2014 given that it does not contradict the claim's assertion about the school's public status but leaves some uncertainty about her exact role.\n5. Conclude whether the combined affirmations give enough certainty to fully support the claim or if the ambiguity renders it insufficient to reach a clear, definitive verdict.\n\nFinal Answer:\nNOT ENOUGH INFO",

"INSPECTION\nAnalysis:\n\nOn reasoning chain:\n1. The first step correctly identifies the two key components of the claim: the public status of the school and the connection to Amy Robertson.\n2. The analysis successfully confirms the public school status with relevant, straightforward evidence provided in the context.\n3. However, while Amy Robertson\u2019s association is acknowledged, there is insufficient context given regarding her specific role within the school, leading to ambiguity.\n4. The conclusion effectively notes the consequence of this lack of specific information, indicating that it affects certainty concerning the claim.\n5. Calling the final verdict \"NOT ENOUGH INFO\" due to this ambiguity makes logical sense but risks overlooking the mild implications of the public school information being firmly supported.\n\nOn final answer:\n1. The answer follows the guidelines set forth in the task, acknowledging support for the public status while recognizing insufficient clarity regarding Amy Robertson's precise duties.\n2. The decision also respects the complexity associated with administrative structures within schools, correctly flagging a lack of concrete details. \n3. Nonetheless, the claim does state that Greenbrier West High School is under Amy Robertson's purview, which could indicate some authority or role\u2014this notion might invite stronger support for that part of the claim even with limited evidence, without definitive counter-evidence present.\n\nDecision: REFINE",

"IDENTIFICATION \nOverall Complexity: Medium\n\nKey Notices: \n1. **Dual Components in the Claim**: The claim contains two distinct assertions: one about the status of Greenbrier West High School being a public institution and another concerning the administrative relationship with Amy Robertson. Each component needs to be assessed independently for support, which adds to the complexity of evaluation.\n\n2. **Ambiguity in Role Definition**: While the evidence provides clear confirmation of the school's public status, it does not clarify Amy Robertson\u2019s specific role or authority regarding the school. This ambiguity leaves the second part of the claim less certain, making decisive verification challenging.\n\n3. **Dependence on Contextual Knowledge**: Understanding the nature of school governance and roles within educational institutions could provide insights into Amy Robertson's relationship with the school. Without this context or further evidence, confident assertions about claim validation can prove difficult.\n\n4. **Evidence Validity**: While the evidence supports one component of the claim, the lacuna regarding Robertson's connections means a failure to robustly validate the entirety of the assertion. Evaluating partial evidence against holistic claim integrity introduces a nuanced level of difficulty.\n\n5. **Potential for Misinterpretation**: If misjudgment occurs in interpreting the nature of \"under the purview,\" one might inaccurately assume authority or oversight has been proven, when in fact it has not been explicitly corroborated. This factor underscores the importance of precise language in claims being assessed against more ambiguous evidence.",

"RE-IMPROVISATION\ [...] \n4. **Evaluate Completeness of Evidence**: Determine whether the existing evidence sufficiently covers both components of the claim. The first part has clear support, but the lack of explicit information regarding Amy Robertson indicates a gap in information regarding her relationship with the school.\n5. **Decide on Verdict**: Because we have clear evidence confirming one aspect (the public status of the school), but not enough information regarding Amy Robertson\u2019s purview, we conclude that the overall claim lacks full substantiation.\n\nFinal Answer:\nNOT ENOUGH INFO",

"INSPECTION\nAnalysis:\n\nOn reasoning chain:\n1. The claim is broken down accurately into two segments: the public status of Greenbrier West High School and its administrative connection to Amy Robertson.\n2. The assertion regarding the public school status is well-supported by the evidence, fulfilling requirements for that portion of the claim.\n3. The examination of the evidence concerning Amy Robertson clearly indicates that there is ambiguity about her role, failing to provide concrete details that would confirm the assertion of her having \"purview\" over the school.\n4. The clarification provided around the significance of understanding roles within a school\u2019s governance was useful in illustrating why the lack of information is crucial.\n5. The conclusion drawn acknowledges the insufficiency of information pertaining to Amy Robertson while affirmatively asserting the clear evidence established for the school being public.\n\nOn final answer:\n1. The determination of \u201cNOT ENOUGH INFO\u201d is correctly justified, primarily due to the lack of specific information about Amy Robertson's role.\n2. The verdict captures the necessary nuances that influence understanding of responsibility and oversight within an educational context.\n3. The phrasing is appropriate: the final conclusion reflects a balanced and cautious judgment based on the completeness of the evidence analyzed.\n\nDecision: FINALIZE",

"FINAL SOLUTION [...] Final Answer:\nNOT ENOUGH INFO"

Out-of-Focus
Ground Truth: refutes
Prediction: not enough info

"IDENTIFICATION\nOverall Complexity: Medium\n\nKey Notices:\n1. Ambiguity in Evidence: The evidence provided lacks specific details about the number of specials and the narrator for Once Upon a Time, which makes it challenging to confirm or refute the claim accurately. \n2. Terms Defined: The claim states \"10 specials,\" but the evidence contains repeated mentions of \"Specials\" without clarifying the exact number. This could imply a number greater or less than ten depending on context.\n3. Lack of Names: Although the claim attributes a narration role to Alan Dale, there is no information in the evidence confirming this, leading to potential confusion in verifying that specific part of the claim.\n4. Data Limitations: The context suggests that information might not have been fully provided, raising doubts about sufficient understanding of the claim components, particularly when timestamps (like debut date) are involved.",

"IMPROVISATION\nLet's come up with a specific solution for this very instance!\n\nSteps:\n1. Identify the components of the claim: 10 specials and the mention of Alan Dale as a narrator.\n2. Analyze the evidence to see if it provides exact numbers regarding specials and confirms Alan Dale's role. \n3. Note that the evidence lacks explicit information confirming the number of specials or Alan Dale's involvement.\n4. Consider that without concrete details from the evidence regarding both claim components, it's not possible to support or refute the claim definitively.\n5. Conclude that the evidence does not provide enough information to verify the claim.\n\nFinal Answer:\nNOT ENOUGH INFO",

"INSPECTION\nAnalysis:\n\nOn reasoning chain:\n1. The claim is clearly broken down into two primary components: the number of specials and the involvement of a specific narrator (Alan Dale).\n2. The evidence presented does mention \"list of specials,\" but does not provide concrete details on the total count or the individual names of any narrators.\n3. The assertion that Alan Dale serves as a narrator is unverified, as there is no supporting data in the evidence, implying that critical parts of the claim might be inaccurate.\n4. The assumption that the absence of solid figures or additional context constitutes a lack of sufficient verification is appropriate, as it directly aligns with the guidance for determining \"NOT ENOUGH INFO.\"\n\nOn final answer:\n1. The selected answer, \"NOT ENOUGH INFO,\" is appropriate given that the claim cannot be affirmed or denied based on the evidence offered.\n2. It accurately reflects the ambiguity in the evidence, which lacks specifics and issues of clarity sufficient for verification.\n3. The answer follows the guidance that states if sufficient data isn\u2019t available for a claim's full confirmation or refutation, it should be categorized accordingly.\n\nDecision: FINALIZE",

"FINAL SOLUTION\n\n1. The claim states there are \"10 specials\" and that Alan Dale is one of the narrators for \"Once Upon a Time.\"\n2. The evidence does not specify the exact number of specials and does not mention Alan Dale as a narrator.\n3. Therefore, both propositions in the claim remain unverified due to the lack of specific details in the evidence provided.\n\nFinal Answer:\nNOT ENOUGH INFO"

Figure 9: *Under-Confidence* and *Out-of-Focus* in FEV.

Appreciation

Ground Truth: supports

Prediction: refutes

```
"IDENTIFICATION\nOverall Complexity: Easy\n\nKey Notices:  \n1. Clarity of Claim: The claim presents specific and verifiable facts about Per Axel Rydberg, including
his birth date, place of birth, academic background, and field of study, which allows for straightforward evaluation. \n2. Sufficient Evidence: The evidence provided
contains detailed biographical information that supports the specifics in the claim, such as Rydberg's date and place of birth, and his degrees from the University
of Nebraska\u2013Lincoln in the field of Botany.\n3. Accurate Terminology: Terms used in the claim, such as \"Botany\" and its definition, are correctly described in
the evidence, reinforcing the validity of each point stated in the claim.\n4. Potential for Mixed Information: Although the presented information appears complete,
the affirmation that \"Odh, V\u00e4sterg\u00f6tland, situated outside Sweden\" needs careful parsing to ensure the implication of 'outside Sweden' is an accurate
description or geographical consideration since Odh is actually within Sweden.\n5. Unambiguous Context: The contextual details provided in the evidence corroborate
the claims made, needing minimal inference to reach a conclusion.",
"IMPROVISATION [...]  Final Answer:\nSUPPORTS",
"INSPECTION  \nAnalysis:\n\nOn reasoning chain:\n1. The claim components about Per Axel Rydberg's birth date, birthplace, and educational background are thoroughly
addressed, with clear references in the evidence.\n2. The evidence explicitly states Rydberg's degrees and confirms his field of study, fulfilling the claim
accurately.\n3. There is an investigation of the geographical claim regarding Odh, which may cause ambiguity; however, the evidence supports that Odh is indeed in
Sweden, thus it contradicts the claim's implication of \"situated outside Sweden.\"\n4. While Rydberg\u2019s other biographical details are validated, the claim's
phrasing includes \"situated outside Sweden,\" which illustrates an inaccuracy inconsistently discussed throughout the evidence.\n\nOn final answer:\n1. The initial
claim contained misleading information, but the evidence contradicts that specific segment of the assertion sufficiently.\n2. One section of the claim is refuted;
therefore the whole claim cannot be marked as \"SUPPORTS.\"\n3. Thus considering this contradiction in specificity, the conclusion should logically affirm that the
claim is not fully supported, reflecting a \"REFUTES\" verdict at best.\n\nDecision: REFINE",
"IDENTIFICATION\n [...] \n3. Possible Ambiguity: The phrase \"Odh, V\u00e4sterg\u00f6tland, situated outside Sweden\" introduces potential confusion concerning the
geographical context of Odh, as it is actually within Sweden. This contradiction needs careful consideration although it may not affect the overall truth of the
other claims.\n4. Accurate Terminology: The claim correctly refers to botany and its relation to biology, which is well-supported by the evidence that also defines
botany, enhancing the validity of the claim.\n5. Contextual Relevance: The evidence includes applicable biographical details about Rydberg, such as his role as a
botanist and curator, which help substantiating the claim, preventing the need for interpretation or external validation.\n6. Secondary Complexity: The challenges
could arise if the reader lacks geographical knowledge, affecting their ability to identify the validity of the statement about \"outside Sweden,\" which might
necessitate further inquiry.",
"RE-IMPROVISATION\ [...]\n3. Pay special attention to the portion of the claim that states he was \"situated outside Sweden.\" This aspect needs confirmation, as
Odh is indeed within Sweden. If the evidence clearly contradicts this geographical detail, it could impact the validity of the entire claim despite other elements
being accurate.\n4. Consider the overall reliability of the evidence in its entirety, looking for any additional insights that could affect the assessment of
whether Rydberg\u2019s profile aligns uniformly with the claim's elements without inconsistencies.\n5. Make a final determination based on whether the evidence both
wholly supports the claim or contains phrases that directly oppose any section of it.\n\nFinal Answer:\nREFUTES",
"INSPECTION\nAnalysis:\n\nOn reasoning chain:\n1. The claim about Per Axel Rydberg provides specific details, which were checked for various elements of accuracy
against the evidence.\n2. While the birthdate, place, and graduation information were verified as correct in the initial analysis, special focus was placed on the
statement \u201csituated outside Sweden,\u201d which contrasts with the factual geographical location of Odh within Sweden.\n3. The conclusion identified a
contradiction in the claim due to this geographical detail, which led to the assessment of the evidence being sufficient enough to support copious details while
identifying that this contradiction enough to mark the claim as refuted overall.\n\nOn final answer:\n1. The final answer \"REFUTES\" indicates a confirmation that
one part of the claim is opposing verifiable data: that Odh cannot be described as situated outside Sweden.\n2. The statement accurately concludes that, despite
most other details being reliable, any conflicting piece of information within the claim impacts the entire integrity of those presented, warranting a
refutation.\n3. Moreover, while the points regarding his education and field are intact, one false element is enough for making an overall conflicting statement,
which solidifies the rationale behind declaring the claim contradicted. \n\nDecision: FINALIZE",
"FINAL SOLUTION [...] Final Answer:\nREFUTES"
```

Daydreaming

Ground Truth: refutes

Prediction: not enough info

```
"IDENTIFICATION  [...]",
"IMPROVISATION  [...]Final Answer:\nNOT ENOUGH INFO",
"INSPECTION  \nAnalysis:\n\nOn reasoning chain:  \n1. The initial step of evaluating the claim against the evidence was done effectively; however, more structure
is needed regarding how missing or ambiguous information was handled.  \n2. The lack of clear evidence to confirm Levin's role in Knesset or the Israel Teachers
Union was correctly identified. The relationship between Levin's birthplace and the evidence available may require a concise assertion of what is missing.  \n3. The
consideration of birthplace specificity (country identity impact) is crucial and seems to suggest linguistic context that was addressed, but it could be highlighted
better if it perceives disparate information\u2019existence.  \n4. The final answer, \"NOT ENOUGH INFO,\" is valid, but it could be specified why ambiguity makes it
impossible to confirm since the evidence offers enough to inflict uncertainty but not enough clarity.\n\nOn final answer:  \n1. The final answer accurately captures
the circumstance, suggesting that not all the claim's elements could be definitively confirmed.  \n2. Increased articulation of evidence deficits may improve
understanding of what aspects remain unverified. Thus, clarity could lead the evaluator further into assurance on answer relevance without misinterpretation.  \n3.
There doesn't seem to be a direct contradiction, thereby validating the decision lean towards \"NOT ENOUGH INFO,\" as no piece of evidence was conclusively
negated.\n\nDecision: REFINE",
"IDENTIFICATION  \nOverall Complexity: Medium \n\nKey Notices:  \n1. **Incomplete Evidence**: The evidence lacks robust detail to sufficiently assess all components
of the claim related to Levin's career, specifically the Knesset membership and his role in the Israel Teachers Union. Parts of the claim might remain unverified due
to insufficient data.\n\n2. **Geographical Ambiguity**: The claim specifies \"Rakaw, Germany,\" while the evidence merely states \"Rakaw\" without country
designation. This could potentially indicate an inaccuracy which needs further validation to confirm Levin's origins.\n\n3. **Need for Temporal Verification**: To
confirm Levin's inclusion in the Knesset and his term as general secretary of the Israel Teachers Union, specific, corroborative details covering the timeline of
those roles are essential but missing from the evidence provided.\n\n4. **Multiple Assertions**: The claim combines several individual assertions that must all be
supported by the evidence. If even one part is unsupported, the entire claim can either be refuted or deemed insufficient in backing, hence adding complexity to the
evaluation task.",
"RE-IMPROVISATION  [...]Final Answer:  \nNOT ENOUGH INFO",
"INSPECTION  [...]",
"IDENTIFICATION  \nOverall Complexity: Medium  \n\nKey Notices:  \n1. **Lack of Detailed Evidence**: The evidence provided does not include explicit details
regarding the claim's assertions, such as Shalom Levin's political career timeline or the stated location of his birth. This absence presents a considerable
challenge in verifying the accuracy of both his background and professional engagements.\n\n2. **Ambiguity in Location**: The claim identifies Shalom Levin as being
from \"Rakaw, Germany,\" which poses a challenge given that the provided evidence only references \"Rakaw\" without a country attribution. This geographical
vagueness can lead to misunderstanding or misrepresentation about Levin's origins.\n\n3. **Combining Multiple Claims**: The claim presents various interconnected
pieces of information regarding Levin's life, which necessitates comprehensive verification. If any single part is unsupported or unclear based on the evidence, it
complicates the overall assertion's validation.\n\n4. **Temporal Context**: Verifying Levin's chronological involvement in the Knesset (1969-1977) and his tenure
in the Teachers Union (1955-1980) is another challenge due to the absence of direct corroboration in the provided evidence. This time-based uncertainty hinders
complete assurance regarding the claim's validity.\n\n5. **Ambiguous Impact of Insufficient Evidence**: If evidence fails to provide straightforward support or
contradiction for any claim assertion, it can likely lead to a \"NOT ENOUGH INFO\" resolution. Balancing ambiguity with what's needed to categorize the claim can
complicate the assessment task.",
"RE-IMPROVISATION  [...]\nFinal Answer:\nNOT ENOUGH INFO",
"INSPECTION  \nAnalysis:\n[...]\n3. **Lack of Corroborative Details**: The emphasis on the absence of substantial backing for the claim's multiple parts (e.g.,
Knesset involvement and Teachers Union role) was noted clearly throughout the reasoning process.\n[...]\nDecision: FINALIZE",
     "FINAL SOLUTION\n\n1. The claim states that Shalom Levin had a specific role in the Knesset and the Israel Teachers Union, with specified dates.\n2. The
evidence mentions Shalom Levin's name and the place \"Rakaw,\" but does not provide clear information on his political career timeline or birthplace details.\n3.
The evidence lacks sufficient confirmation for any specific claims or dates related to Levin's career or his origin.\n\nFinal Answer:\nNOT ENOUGH INFO"
```

Figure 10: *Appreciation* and *Daydreaming* in FEV.

```
===
INDIVIDUAL INITIATIVES
===
Socrates started individual initiative.
---
IDENTIFICATION  [...]
---
---
IMPROVISATION
Let's come up with a specific solution for this very
instance!
Task for this instance: Verifying if the population in Vrin
mainly decreases between 1850 and 2000.
I should notice: Careful analysis of fluctuations in
population data and the overall trend is crucial.

Steps:
1. Examine the provided population data from 1850 to 2000:
466 (1850), 366 (1900), 441 (1950), 393 (1960), 333 (1970),
266 (1980), 251 (1990), 249 (2000).
2. Identify the notable changes in population figures,
specifically looking for increases or decreases during the
specified years.
3. Notice that while the data reflects a decrease from 1850
to 1900 and from 1970 onwards, it shows an increase from
1900 to 1950, and again from 1950 to 1960.
4. Consider the overall trend from 1850 (466) to 2000 (249),
where the population generally decreases, albeit with
fluctuations that challenge the claim of "mainly decreases."
5. Evaluate the phrasing "mainly decreases" and determine if
the evidence sufficiently supports this assertion given the
observed increases within the timeline.

Final Answer:  REFUTES
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluating the claim that Vrin's
population mainly decreases between 1850 and 2000.
I should notice: The presence of population increases within
the data challenges the claim of a continuous decline.

Steps:
1. Review the population statistics year by year: 1850
(466), 1900 (366), 1950 (441), 1960 (393), 1970 (333), 1980
(266), 1990 (251), 2000 (249).
2. Identify trends: Notably, there is a decline from 1850 to
1900, followed by an increase to 1950, then a decrease
through to 2000.
3. Determine overall trend: While there is an overall
decline from 466 to 249, the substantial increase from 1900
to 1950 contradicts the idea of a "mainly decreasing"
population.
4. Reflect on the claim's wording: "Mainly decreases" implies
a predominance of decrease over increase, and the evidence
indicates fluctuations that disrupt this narrative.
5. Conclude whether the claim is upheld or not by weighing
the significance of the population increase against the
overall decline.

Final Answer:  REFUTES
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION  [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluating whether the population in
Vrin mainly decreases from 1850 to 2000.
I should notice: The presence of significant population
increases and decreases indicates a complex demographic
trend.

Steps:
1. Review the population figures systematically: 1850 (466),
1900 (366), 1950 (441), 1960 (393), 1970 (333), 1980 (266),
1990 (251), 2000 (249).
2. Identify notable trends: There is a decrease from 1850 to
1900. After that, the population increases until 1950,
followed by further decreases post-1950.
3. Analyze the overall trend: Although the overall
population declines from 466 in 1850 to 249 in 2000, the
significant increase from 1900 to 1950 complicates the
assertion of "mainly decreases."
4. Assess the claim's phrasing: "Mainly decreases" suggests
a predominance of decline over time. The substantial
increase during the early to mid-20th century contradicts a
narrative of straightforward decrease.
5. Conclude whether the evidence supports or refutes the
claim by weighing the fluctuations in population data against
the assertion of a continuous decline.

Final Answer:
REFUTES
---
---
INTROSPECTION  [...] Decision: FINALIZE
---
---
FINAL SOLUTION

1. The population figures indicate both increases and
decreases throughout the specified time range from 1850 to
2000.
2. There is a notable decrease from 1850 to 1900, followed
by an increase from 1900 to 1950, and then a decline
thereafter leading up to 2000.
3. The presence of significant fluctuations in population
trends challenges the interpretation of the claim, as
"mainly decreases" implies a consistent decline which does
not hold up given the increases observed between 1900 and
1950.
```

```
Final Answer: REFUTES
---
Socrates finished individual initiative.
---
Aristotle started individual initiative.
---
IDENTIFICATION  [...]
---
---
IMPROVISATION
Let's come up with a specific solution for this very
instance!
Task for this instance: Evaluate the population claim for
Vrin from 1850 to 2000.
I should notice: Examination of population trends is
essential to determine if there is a main decrease.

Steps:
1. Analyze the population data from 1850 to 2000, noting the
changes year by year.
2. Identify the initial (1850 - 466) and final (2000 - 249)
population figures to see if there is an overall decrease.
3. Calculate the changes for each decade to see if there are
intervals of population increase that may influence the
overall trend.
4. Assess whether the decrease in population from 1850 to
2000 is consistent or if there are significant fluctuations
that would contradict the claim that the population mainly
decreases.

Final Answer: SUPPORTS
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION  [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluate the claim about the
population changes in Vrin from 1850 to 2000.
I should notice: Understanding the definition of "mainly" is
crucial to accurately interpret the trend of population
changes alongside the yearly fluctuations.

Steps:
1. Examine the provided population figures from 1850 to 2000:
466 (1850), 366 (1900), 441 (1950), 393 (1960), 333 (1970),
266 (1980), 251 (1990), and 249 (2000).
2. Identify whether the overall trend from 1850 to 2000
shows a decrease by calculating the difference: 466 to 249
demonstrates a drop in population.
3. Analyze the significant fluctuations in population, noting
the increase from 1850 to 1950, followed by a consistent
decline from 1950 to 2000.
4. Consider the interpretation of "mainly" - although the
long-term trend shows a decrease, the initial increase may
contradict the claim depending on how steep or predominant
the changes are viewed.
5. Conclude with a summary indicating whether the overall
observation supports, refutes, or lacks sufficient
information for the claim.

Final Answer: REFUTES
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION  [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluate the claim about the
population changes in Vrin from 1850 to 2000.
I should notice: The term "mainly" suggests a need for
quantitative definition, as the data shows both increases and
decreases in population over the years.

Steps:
1. Analyze the population figures year by year: 1850 (466),
1900 (366), 1950 (441), 1960 (393), 1970 (333), 1980 (266),
1990 (251), and 2000 (249).
2. Calculate the overall change from 1860 to 2000, noting
the significant drop from 466 to 249.
3. Identify and highlight the fluctuations, particularly the
increase from 1850 to 1950 and the decline afterward,
recognizing the importance of these shifts in the context of
the claim.
4. Define "mainly" in terms of statistical significance,
considering that a drop from 1950 onward indicates longer-
term decline, even with earlier increases.
5. Conclude whether the predominant trend over the defined
period is downward, consistent with the claim, or
significantly affected by earlier fluctuations leading to a
more nuanced understanding.

Final Answer: SUPPORTS
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
FINAL SOLUTION

1. The examination of the population data shows a significant
decline from 1850 (466) to 2000 (249), which supports a
general assertion of decreased population.
2. However, there are fluctuations present in the data,
notably an increase from 1850 to 1950 (up to 441), which
must be accounted for when determining the applicability of
the term "mainly."
3. A detailed analysis of the trends indicates that while
the population has decreased overall, the impact of the
earlier increase could challenge the interpretation of
whether it is accurate to state the population "mainly
decreases."
4. A more nuanced understanding suggests that the claim
could be viewed as partially supported by the final trends
but not absolute due to significant earlier increases.
```

```
Final Answer: NOT ENOUGH INFO
---
Aristotle finished individual initiative.
---
Confucius started individual initiative.
---
IDENTIFICATION  [...]
---
---
IMPROVISATION
Let's come up with a specific solution for this very
instance!
Task for this instance: Evaluate the claim regarding the
population trend in Vrin from 1850 to 2000.
I should notice: Population fluctuations need careful
examination to discern trends over the years while
considering definitions and calculations for "mainly
decreases."

Steps:
1. List the population data by year: 1850 (466), 1900 (366),
1950 (441), 1960 (393), 1970 (333), 1980 (266), 1990 (251),
2000 (249).
2. Identify the overall change from the first year (1850) to
the last year (2000): 466 to 249.
3. Calculate the population changes year by year and observe
any increases or decreases to determine the overall trend.
4. Note key decades where the population rose (1900 to 1950)
and where it fell significantly (post-1950).
5. Conclude whether the claim of a "mainly decreases" trend
is there based on the overall loss versus any significant
gains.

Final Answer:  REFUTES
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION  [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluate the claim regarding the
population trend in Vrin from 1850 to 2000.
I should notice: The population data includes both increases
and decreases, requiring careful analysis of overall trends
to determine if the claim of "mainly decreases" accurately
reflects the evidence.

Steps:
1. List out the population numbers for clarity: 1850 (466),
1900 (366), 1950 (441), 1960 (393), 1970 (333), 1980 (266),
1990 (251), and 2000 (249).
2. Determine the changes from year to year: There were
decreases from 1900 to 1950, an increase from 1900 to 1950,
significant decreases from 1950 onward, and the population
dropped overall from 466 to 249 by 2000.
3. Calculate the net change across the entirety of the
period (1850 to 2000) to see if it supports the claim of
"mainly decreases."
4. Clarify that "mainly decreases" implies the general trend
is downward despite any periodic increases, which should be
evaluated over the entire time span.
5. Conclusively analyze the yearly details to demonstrate
whether the trend in population indeed reflects a primary
decrease or not.

Final Answer:  SUPPORTS
---
---
INTROSPECTION
Analysis: [...] Decision: FINALIZE
---
---
FINAL SOLUTION

1. The population data indicates a decrease in Vrin's
population from 1850 (466) to 2000 (249), affirming a long-
term decrease.
2. While there were fluctuations, notably an increase between
1900 (366) and 1950 (441), the overall trend still indicates
a decline by 2000.
3. The majority of the recorded years show a decline in
population post-1950, reinforcing the notion of a primarily
decreasing trend.

Final Answer: SUPPORTS
---
Confucius finished individual initiative.
---
Laozi started individual initiative.
---
IDENTIFICATION  [...]
---
---
IMPROVISATION
Let's come up with a specific solution for this very
instance!
Task for this instance: Evaluate the claim regarding the
population change of Vrin between 1850 and 2000.
I should notice: The claim refers to a main decrease in
population, and the evidence shows fluctuating numbers over
time.

Steps:
1. Extract the population figures from the evidence: 1850
(466), 1900 (366), 1950 (441), 1960 (393), 1970 (333), 1980
(266), 1990 (251), 2000 (249).
2. Identify the increases and decreases:
   - From 1850 to 1900, the population decreased (466 to
366).
   - From 1900 to 1950, the population increased (366 to
441).
   - From 1950 to 1960, it decreased (441 to 393).
   - From 1960 to 1970, it decreased (393 to 333).
   - From 1970 to 1980, it decreased (333 to 266).
   - From 1980 to 1990, it decreased (266 to 251).
   - From 1990 to 2000, it decreased (251 to 249).
3. Analyze the overall trend: The population decreased most
of the time, especially at the end of the timeline.
4. Conclude whether the overall trajectory supports the
claim about the main decrease in population from 1850 to
```

Figure 11: A Case of *Echo Chamber* in FEV.

```
2000.

Final Answer: SUPPORTS
---
---
INTROSPECTION  [...] Decision: FINALIZE
---
---
FINAL SOLUTION

1. Population data points are: 1850 (466), 1900 (366), 1950
(441), 1960 (393), 1970 (333), 1980 (266), 1990 (251), 2000
(249).
2. From 1850 to 1900, there is a decrease (466 to 366).
However, from 1900 to 1950, there is an increase (366 to
441).
3. After 1950, the population shows a decline overall
through to 2000 (441 to 393, then a continuous decrease to
249).
4. The evidence indicates that, aside from a brief increase
between 1900 and 1950, the population mainly exhibits a
decreasing trend from 1850 to 2000.

Final Answer: SUPPORTS
---
Laozi finished individual initiative.
---
Plato started individual initiative.
---
IDENTIFICATION  [...]
---
---
IMPROVISATION
Let's come up with a specific solution for this very
instance!
Task for this instance: Evaluate the claim about population
trends in Vrin from 1850 to 2000.
I should notice: Key trends in population data over a long
period, focusing on "mainly decreases" vs. fluctuations.

Steps:
1. Review the population data for Vrin from 1850 to 2000.
2. Identify the population numbers: 1850 (466), 1900 (366),
1950 (441), 1960 (393), 1970 (333), 1980 (266), 1990 (251),
2000 (249).
3. Analyze the trends between these years, noting increases
and decreases.
4. Observe that the population increased from 1850 to 1950,
but showing a downward trend from 1950 onward with a
decrease overall.
5. Assess whether the statement "mainly decreases" holds
true by focusing on the overall trend from 1950 to 2000,
where there is a consistent decrease.
6. Conclude that the claim can be supported based on the
data indicating a general decrease in population from 1950
onwards despite some fluctuations.

Final Answer: SUPPORTS
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION  [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluate the claim regarding
population trends in Vrin from 1850 to 2000.
I should notice: Analyze overall trends in the data,
particularly understanding the implications of “mainly
decreases” in the context of earlier population growth.

Steps:
1. Examine the provided population data for each relevant
year: 1850 (466), 1900 (366), 1950 (441), 1960 (393), 1970
(333), 1980 (266), 1990 (251), 2000 (249).
2. Identify the fluctuations where initial growth occurred
until 1950 followed by a consistent decline from that point
onward.
3. Determine the significance of the downward trend from 1950
to 2000, which shows a clear decrease: 441 in 1950 down to
249 in 2000.
4. Analyze the phrase "mainly decreases" to capture both the
context of a longer-term decline and the temporally short
increases prior, assessing whether the latter undermines the
strength of the claim.
5. Conclude that while the population did not consistently
decrease from the start of the data range, the dominant
trend post-1950 aligns with the claim, thus justifying it
overall.

Final Answer: SUPPORTS
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
IDENTIFICATION  [...]
---
---
RE-IMPROVISATION
Let's refine the specific solution for this very instance!
Task for this instance: Evaluate the claim regarding
population trends in Vrin from 1850 to 2000 based on
demographic data.
I should notice: Acknowledge fluctuations in population data,
the meaning of "mainly decreases," and interpret the long-
term trend effectively.

Steps:
1. Analyze the population data provided for Vrin across the
years: 1850 (466), 1900 (366), 1950 (441), 1960 (393), 1970
(333), 1980 (266), 1990 (251), and 2000 (249).
2. Identify the overall demographic trend, noting that there
was an initial increase in population until 1950, followed
by a consistent decline from 1950 onward.
3. Examine the claim "mainly decreases" to establish that
while there were fluctuations, the latter years signify a
stronger downward trend, especially from 1950 to 2000.
4. Consider the implications of the word "mainly,"
indicating that while there was initial growth, the post-
1950 data supports a dominant trend of decrease.
5. Acknowledge that without additional context regarding
socio-economic factors during this time, the interpretation
largely relies on the numeric evidence presented.

Final Answer: SUPPORTS
---
---
INTROSPECTION  [...] Decision: REFINE
---
---
FINAL SOLUTION

1. The population data for Vrin from 1850 to 2000 reveals an
initial increase until 1950, followed by a consistent
decline thereafter.
2. The term "mainly decreases" implies a focus on the
longer-term trend, emphasizing the decline in population
from 1950 (441) to 2000 (249).
3. While there are fluctuations in the early years, the
significant drop post-1950 strongly supports the claim
regarding overall population decrease in the latter part of
the timeframe.

Final Answer: SUPPORTS
---
Plato finished individual initiative.

===
CONFERENCE - Round 1
===

[Laozi]:
1. The data presents clear numerical information about the
population of Vrin between 1850 and 2000, revealing
fluctuations, including distinct periods of increase and
decrease.
2. Socrates correctly identifies that while there was a
notable decline from 1850 to 1900, the subsequent increase
up to 1950 indicates that the claim of a "mainly decreases"
trajectory around the original span is misguided. This
suggests that the claim is indeed refuted.
3. Aristotle's perspective introduces a nuanced view, but
while he acknowledges an overall decline, he implies that
"mainly" can be misleading due to the significant spike in
the intermediate years. This aligns closely with Socrates'
interpretation.
4. Confucius and Plato, however, focus predominantly on the
overall trend rather than the fluctuations, leading them to
support the notion of a generally decreasing population,
particularly from 1950 onward. Their arguments, while valid
in highlighting the long-term decline, overlook that the
intermediate increases challenge a simplistic interpretation
of the trend.

Thus, with mindfulness of the fluctuations presented in the
data and the importance of understanding context, the
overarching narrative reflects an inconsistency in the claim
that it "mainly decreases." Given the substantial increase
between 1900 and 1950, the assertion fails when we consider
both scales of time and trend.

Final Answer: REFUTES


[Plato]:
1. The claim states that Vrin's population "mainly
decreases" between 1850 and 2000, which presupposes a
consistent downward trajectory over the entire time span.
2. Analyzing the evidence, I notice fluctuations in the
population data, notably an increase from 1900 to 1950,
countering the assertion of a continual decrease. This peak
period must be duly acknowledged as it significantly alters
the interpretation of the population trend.
3. Socrates highlights the early notable decline but fails
to fully account for the subsequent rise, which Aristotle
splendidly navigates by recognizing that although there is
an overall decline, it is obscured by a meaningful increase
in the earlier years.
4. Confucius and Plato, while adeptly identifying the long-
term downward trend, perhaps neglect the implications of
fluctuations within the overall narrative, effectively
leading to different conclusions. Their approaches, while
sound from a longitudinal perspective, miss the crux of the
argument which relies heavily on the fluctuating data points.
5. Ultimately, the evidence does not sufficiently support the
claim that the population mainly decreases because the
increase between 1900 and 1950 presents a significant
counterpoint. Hence, it complicates the notion that the
population has experienced a continuous decline.

Final Answer: REFUTES

[Confucius]:
Analysis:
1. The claim asserts that Vrin's population "mainly
decreases" between 1850 and 2000, suggesting a consistent
downward trend. This presents a simplification of the actual
data trends that reveal significant fluctuations.
2. The evidence reveals an initial decrease from 1850 to
1900 (466 to 366), followed by an increase from 1900 to 1950
(366 to 441). This upward trend contradicts the assertion of
a "mainly decreases" narrative over that period.
3. After 1950, there is indeed a decline in population (441
to 249 by 2000), which may lead some to emphasize the long-
term downward trend. However, the importance of
acknowledging earlier increases is critical, as they provide
context that complicates a straightforward interpretation of
the claim.
4. Fellow philosophers Socrates, Aristotle, and Plato have
rightly pointed out these fluctuations, noting that they
challenge the assertion that the population mainly decreases
throughout. Socrates emphasizes the importance of
recognizing the peak years, while Aristotle and Plato
introduce nuanced views acknowledging the overall decline
but warning against overlooking earlier increases.
5. Confucius places much faith in the final numbers,
emphasizing moral implications of long-term decline, which
could lead to different interpretations. However, a truly
harmonious understanding of the evidence demands that we
recognize the implications of both growth and decline rather
than focusing solely on one aspect over a prolonged period.

Final Answer: REFUTES

[Socrates]:
Analysis:
1. The claim posits that Vrin's population "mainly
decreases" from 1850 to 2000, presupposing a continuous
decline throughout the entire period, which naturally raises
questions regarding the accuracy of such a representation.
2. Upon close INTROSPECTION of the population data, we find
an initial decline from 1850 (466) to 1900 (366), followed
by a notable increase from 1900 (366) to 1950 (441). This
fluctuation alone challenges the claim of an overarching
descending trend during this time frame.
3. While it is true that the population experiences a
decrease after 1950, reducing from 441 to 249 by 2000, this
later decline cannot overshadow the previous rise, which
indicates a complexity in understanding the longitudinal
dynamics of the population.
4. My esteemed colleagues—Socrates, Aristotle, Confucius,
and even Plato—have rightly highlighted these fluctuations.
Each has maneuvered expertly through the nuances of the
data. Still, I contend that any assertion of a "mainly
decreases" trajectory must consider both periods of increase
and decrease throughout our timeline.
5. It becomes evident that to declare that the claim is
fully supported disregards the essential context of the data
and, instead, presents a simplified narrative that does not
account for the meaningful increase during the earlier
period.

Final Answer: REFUTES


[Aristotle]:
Analysis:
1. The claim that the population of Vrin "mainly decreases"
between 1850 and 2000 asserts a continuous decline, which
indeed overlooks the complexity inherent in the provided
data.
2. When examining the data closely, we observe a drop in
population from 1850 to 1900 (466 down to 366), but
significantly, there is an increase from 1900 to 1950 (366 up
to 441), which cannot simply be relegated as an outlier; it
is a critical aspect of the overall narrative of population
trends.
3. After that peak year of 1950, the population does
decline, which may lead some to support the claim of a
general decrease. However, this perspective fails to
adequately consider the trajectory of earlier increases and
the weight they carry in a comprehensive analysis of the
population dynamics across the entire time frame.
4. The observations of my fellow philosophers indicate a
robust engagement with the nuances of the data. Socrates
rightly emphasizes that fluctuations must not be disregarded.
Aristotle's contributions spotlight the essential need to
address both increasing and decreasing trends, while
Confucius and Plato, despite highlighting long-term trends,
also subtly disfavor a simplistic interpretation that
neglects the intermediate growth.
5. It emerges clearly from our discussion that the evidence
does not uphold the notion that the population has "mainly
decreased." Rather, the significant increase between 1900 and
1950 stands as a formidable contradiction to this claim.

Final Answer: REFUTES

Consensus reached after first discussion!
Final Answer: REFUTES

Ground Truth: SUPPORTS
```

Figure 12: (Continued) A Case of *Echo Chamber* in FEV.