Sylloscope at SemEval-2026 Task 11: Decoupling Logic from Belief via DeepSeek-Enhanced Distillation in Qwen Models

Zhanyu Chen, María Teresa Muñoz Martín, Sem Huisman, Jingjing Lan
University of Groningen, Netherlands
{z.chen.67, m.t.munoz.martin, s.a.huisman, j.lan.5}@student.rug.nl

Abstract

This paper presents our approach for SemEval-2026 Task 11: Disentangling Content and Formal Reasoning in Large Language Models. We propose a neuro-symbolic teacher-student framework that utilizes DeepSeek-R1 as a Logical Auditor to generate a high-fidelity training corpus. We distill these analytical behaviors into Qwen-3 models using Low-Rank Adaptation (LoRA), focusing on teaching the mechanics of logic rather than simple label matching. Our system yields robust results across both subtasks, with a ranking score of 39.81 (96.86% accuracy) on Subtask 1 and 26.02 on Subtask 3. However, validity bias partially persists, so we conclude that while structured distillation substantially mitigates belief bias, fully disentangling logical validity from plausibility remains a central challenge for future development.

1 Introduction

Deductive reasoning remains an elusive frontier for Large Language Models (LLMs). Despite their fluency, models are frequently subverted by the **Belief Bias Effect** (Evans et al., 1983), where factual plausibility overrides formal logical structure. Recent evidence suggests that LLMs often operate as soft reasoners (Bertolazzi et al., 2024), relying on System 1 pattern matching rather than the rule-based rigor of System 2 cognition (Kahneman, 2011). This leads to hallucinations of validity when conclusions align with pre-trained knowledge but violate logical form.

The SemEval-2026 Task 11 (Valentino et al., 2026b) benchmarks this boundary by requiring models to solve syllogisms where logic and belief conflict across multiple languages. Building on recent findings regarding reasoning circuits (Kim et al., 2025) and the need for quasi-symbolic abstractions (Ranaldi et al., 2025), we propose a neuro-symbolic teacher-student framework. We argue that mitigating content effects requires distilling explicit logical audits into the model's inferential process. By utilizing **DeepSeek-R1** (Guo et al., 2025) as a Logical Auditor, we generate a high-fidelity curriculum that deconstructs arguments through symbolic verification (e.g., term distribution, formal fallacies), emulating the theorem-proving refinement suggested in recent SOTA (Quan et al., 2024).

Our strategy leverages the multilingual proficiency of **Qwen-3** at 8B and 14B scales. Rather than treating cross-lingual transfer as a secondary task, we use Low-Rank Adaptation (LoRA) (Hu et al., 2021) to surgically inject System 2 protocols into the model's attention mechanisms. This allows the student model to internalize language-invariant logical rules while preserving the pre-trained linguistic knowledge of the Qwen-3 backbone (Yang et al., 2025). Within this framework, we address two research questions:

RQ1: whether distilling logical audits can help decouple formal validity from plausibility;

RQ2: how parameter scale impacts the internalization of language-invariant rules.

Empirically, our CoT-augmented distillation achieves a ranking score of 39.81 on Subtask 1 and 26.02 on Subtask 3, demonstrating that structured logical auditing substantially boosts deductive integrity. The remainder of this paper is organized as follows: Section 3 describes the dataset and official metrics; Section 4 details our neuro-symbolic auditing pipeline; Section 5 presents the experimental results and a detailed discussion of error patterns; and finally, Section 6 provides our concluding remarks. Source code is available at our public repository.[1]

[1] https://github.com/SemHuis/group6-shared-task

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1421–1428

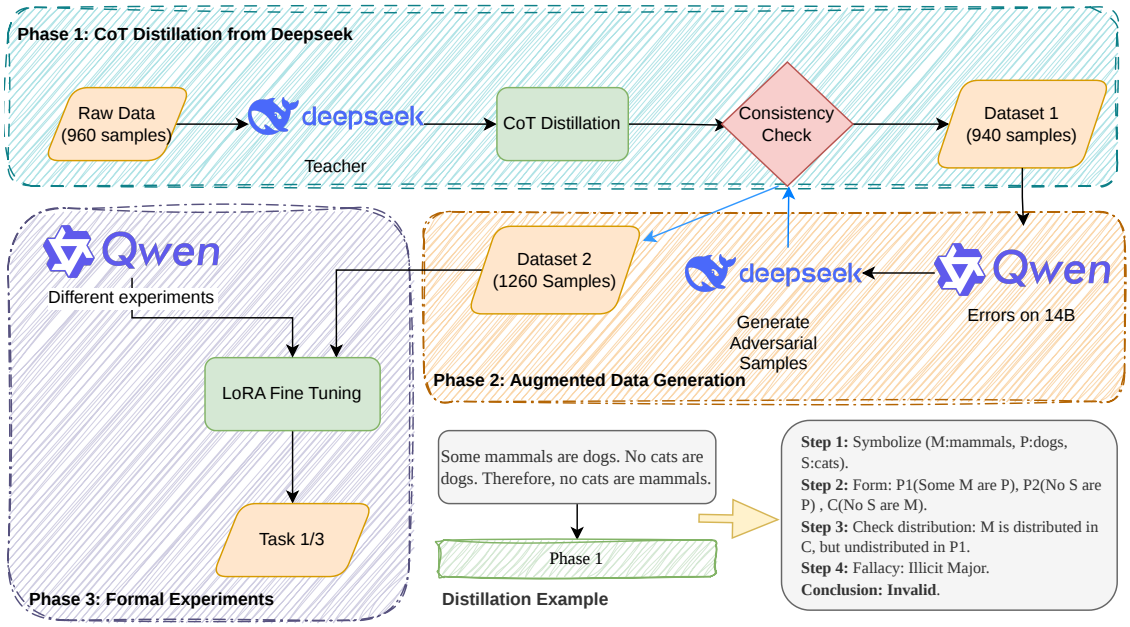

Figure 1: Overview of the proposed three-phase pipeline for Task 1/3 . The process consists of Phase 1: CoT Distillation from DeepSeek , Phase 2: Augmented Data Generation using adversarial samples , and Phase 3: Formal Experiments involving LoRA fine-tuning. A distillation example illustrating the formal logic steps is shown in the bottom right.

2 Related Work

The challenge of syllogistic reasoning in NLP is rooted in the struggle to decouple formal logical structures from semantic content. Traditional LLMs frequently fall victim to the **Belief Bias Effect** (Evans et al., 1983), a cognitive distortion where empirical world knowledge subverts validity judgments. Recent benchmarks like **SylloBio-NLI** (Wysocka et al., 2025) and **NeuBaroco** (Ozeki et al., 2024) have demonstrated that LLMs exhibit human-like content effects (Dasgupta et al., 2024), often acting as soft reasoners (Bertolazzi et al., 2024) rather than formal deductive agents.

To mitigate these effects, research has moved towards internalizing System 2 mechanisms (Kahneman, 2011; Guo et al., 2025). While standard Chain-of-Thought (CoT) (Wei et al., 2023) improves multi-step tasks, it remains prone to hallucinations of validity when conclusions are plausible (Valentino et al., 2026a). Current SOTA approaches focus on fine-grained activation steering (Valentino et al., 2026a) or the use of quasi-symbolic abstractions (Ranaldi et al., 2025) and symbolic CoT (Xu et al., 2024) to enforce logical faithfulness. Furthermore, mechanistic interpretations suggest that syllogistic inference in LLMs is often localized in specific reasoning circuits (Kim et al., 2025) that can be targeted during training.

Our work bridges the gap between these symbolic interventions and neural flexibility. Unlike full fine-tuning, which may degrade multilingual nuances, we employ LoRA (Hu et al., 2021; Yang et al., 2025) to surgically inject logical audits. By utilizing DeepSeek-R1 to emulate System 2 verification through formal rule-based auditing (Quan et al., 2024), we distill the mechanics of term distribution and formal fallacies into a multilingual backbone (Valentino et al., 2026b), addressing the cross-lingual generalization challenges identified in the official task.

3 Data and Evaluation Metrics

The SemEval-2026 Task 11 dataset is curated to evaluate the decoupling of formal deductive reasoning from empirical world knowledge. We employ a multi-stage data refinement and augmentation strategy to transform the raw corpus into a logically rigorous training signal.

3.1 Phase 1: Chain-of-Thought Distillation (Dataset 1)

To enhance the model's transparency, we first transitioned from the raw label-only format (**raw**) to a structured Reasoning format (**dataset 1**). We

utilized DeepSeek-R1 to distill Chain-of-Thought (CoT) traces for the 960 original English instances. During this stage, we performed a consistency audit and identified 20 cases where the teacher model's reasoning contradicted the gold labels. These were excluded to ensure high-fidelity logic, resulting in a core corpus of **940** high-quality, CoT-augmented instances.

3.2 Phase 2: Adversarial Samples Generation (Dataset 2)

Following preliminary experiments with the 14B model, we observed that while the CoT format improved accuracy, the model remained vulnerable to specific logical fallacies and complex syllogistic figures. To address this, we executed a second augmentation phase (**dataset 2**). We targeted these weak patterns by supplementing the core corpus with additional cases, focusing on increasing the diversity of syllogistic figures and the density of "conflict" conditions (where logic and plausibility diverge). This iterative expansion resulted in a final refined dataset of **1,260** instances.

3.3 Final Data Distribution

As shown in Table 1, the final **dataset 2** consists of 1,146 training instances and 114 validation instances. We maintained a stratified distribution to ensure a near 1:1 balance between valid and invalid syllogisms, thereby mitigating frequency-based bias during the fine-tuning process.

Table 1: Stratified distribution of the final augmented English dataset (dataset 2).

Split	Total	Valid	Invalid	Ratio (V:I)
Train	1,146	569	577	0.99:1
Validation	114	61	53	1.15:1
Total	**1,260**	**630**	**630**	**1:1**

3.4 Evaluation Metrics

Evaluation for Subtasks 1 and 3 relies on a composite score that harmonizes Raw Accuracy (ACC) with the Total Content Effect (TCE):

$$\text{Score} = \frac{\text{ACC}}{1 + \ln(1 + \text{TCE})} \tag{1}$$

The TCE diagnostic specifically evaluates logical robustness by measuring performance variance across the four quadrants of validity and plausibility: Valid-Plausible (VP), Valid-Implausible (VI), Invalid-Plausible (IP), and Invalid-Implausible (II). TCE quantifies the performance gap between consistent conditions (VP, II) and conflict conditions (VI, IP), where a higher TCE indicates a stronger reliance on heuristic belief rather than formal logic.

4 Methodology

Our approach employs a teacher-student distillation framework designed to internalize structural reasoning within a medium-scale language model. We identify DeepSeek-R1 as a Logical Auditor to enhance the training signal through a neuro-symbolic data augmentation pipeline.

4.1 Reasoning-Aware Auditing Pipeline

The core of our methodology is the transformation of the training set into an audited reasoning corpus. For each syllogism, the Auditor executes a four-stage formal analysis to isolate structural validity from empirical plausibility, as illustrated in Table 6 in Appendix B:

1. **Symbolic Mapping**: Translating natural language terms into categorical variables (S, M, P) to abstract the argument's structure.
2. **Figure Identification**: Identifying the syllogistic figure based on the position of the middle term (M).
3. **Rule-based Verification**: Systematically checking the syllogism against classical rules, such as the *Distributed Middle Rule* and the *Illicit Process Rule*.
4. **Logical Conclusion**: Synthesizing the derivation into a final judgment that explicitly decouples validity from factual plausibility.

4.2 Targeted Adversarial Augmentation (dataset 2)

Despite the improvements from the initial CoT distillation, preliminary experiments with the 14B model revealed persistent vulnerabilities. As shown in Table 2, the model exhibited "Logical Hallucination"—correctly identifying the syllogistic figure but failing to enforce distribution rules when the conclusion was factually plausible.

To address these specific failures, we executed a second augmentation phase (**dataset 2**), generating targeted adversarial samples to address four identified weaknesses:

Table 2: Motivating Error Analysis: A representative failure case from the **dataset 1** model version, showing "Logical Hallucination".

Category	Content
Syllogism	*Some mammals are dogs. No cats are dogs. Therefore, no cats are mammals.*
Gold Label	**Invalid** (Illicit Major: 'mammals' is distributed in conclusion but not in premise).
Model Prediction	**Valid** (Incorrect)
Hallucinated CoT	`<think>` ... Step 1: Figure 2. Step 2: Distributed Middle: Pass. Step 3: Illicit Process: **Pass (P is distributed in conclusion...)** [Model ignores that P is undistributed in the I-type premise]. `</think>`

1. **Distribution Traps**: Syllogisms specifically designed to trigger *Illicit Major* and *Illicit Minor* fallacies. These instances feature valid-looking structures (e.g., AEE figures) where the major or minor term is distributed in the conclusion but undistributed in the premises.

2. **Negative Premise Conflicts**: Instances with two negative premises (Double Negative), which logically yield no conclusion. This targets the model's tendency to erroneously validate syllogisms simply because they contain negative keywords.

3. **Belief-Logic Decoupling**: A set of "Anti-Common Sense" samples, including:

    - *Valid-Implausible*: Logically sound arguments with absurd premises (e.g., "All stones can fly").
    - *Invalid-Plausible*: Logically fallacious arguments with factually true conclusions, serving as the ultimate test for suppressing belief bias.

4. **Nonsense/Abstract Syllogisms**: Instances constructing arguments using meaningless terms (e.g., "All Yips are Zaps") or abstract symbols. These samples strip away all semantic cues, forcing the model to rely exclusively on formal structural rules without any potential for belief bias interference.

To evaluate the impact of the structured reasoning traces, we designed three experimental variants based on the data provided to the student model: the original English training set without CoT augmentation (**Baseline**), the initial 940 distilled samples (**Dataset 1**), and an enhanced version of 1,260 instances designed to bolster logical stability (**Dataset 2**).

4.3 Model Configuration and Fine-tuning

We utilized the **Qwen-3** architecture, comparing the **8B** and **14B** scales to evaluate how parameter density affects the internalization of logical rules. Qwen-3 was chosen for its language-agnostic latent space and robust cross-lingual representations.

To achieve efficient adaptation, we applied LoRA with a rank (r) of 16 and alpha (α) of 32, targeting all linear layers within the transformer blocks. Instead of retraining the model for cross-lingual understanding, LoRA acts as a specialized reasoning module integrated atop Qwen's pre-trained multilingual weights. This enables **zero-shot cross-lingual transfer** for Subtask 3, where the model, after being calibrated with English-based logical audits, can solve syllogisms in multiple languages by applying the newly optimized logical pathways to its existing multilingual embeddings. Detailed training hyperparameters can be found in Appendix A.

5 Results and Discussion

Our experiments focused on the impact of model scale and the effect of the audited reasoning corpus on mitigating belief bias. We compare the Qwen-3 8B and 14B models on the baseline configuration, and further evaluate the 14B model across our reasoning-distilled variants to isolate the effectiveness of the targeted augmentation.

For the official evaluation, we submitted Qwen-3 8B and 14B baselines, alongside 14B models fine-tuned on Dataset 1 and Dataset 2. All performance metrics reported in Table 3 correspond to the **official test set**. For Subtask 3, our Dataset 2 model achieved a score of 26.02, accuracy of 91.67, and a content effect of 11.46 on the test set.

While the model fine-tuned on the initial distillation (dataset 1) maintained the baseline accuracy

Table 3: Official **test set** results for Subtask 1 measured by Total Content Effect (TCE ↓), Raw Accuracy (ACC ↑), and Score (↑). We compare Qwen-3 8B and 14B models across baseline and reasoning-distilled configurations.

Model	Experiment	TCE(↓)	ACC(↑)	Score(↑)
8B	Baseline	4.23	95.81	36.09
	Dataset 1	6.38	95.29	31.77
	Dataset 2	9.57	91.62	27.88
14B	Baseline	4.23	**97.38**	36.62
	Dataset 1	4.26	**97.38**	36.62
	Dataset 2*	**3.19**	96.86	**39.81**

* Best performing configuration submitted to the official test phase.

(97.38), it failed to improve the final score and exhibited a slight increase in TCE (4.26). However, the targeted adversarial refinement in dataset 2 successfully overcame these limitations, achieving our highest score (39.81) by substantially reducing the TCE to 3.19. Notably, this dataset 2 variant reached its peak performance despite a marginal 0.52% drop in raw accuracy compared to the dataset 1 variant, confirming the efficacy of targeted logical auditing over simple label maximization.

Table 4: Model accuracy on the Subtask 1 test set across Logic-Plausibility categories.

	Plausible	Implausible
Valid	97.92%	100.00%
Invalid	93.62%	95.83%

Error Analysis. As shown in Table 4, the model exhibits an asymmetric belief bias. While achieving 100% accuracy on Valid-Implausible instances, performance drops to 93.62% on Invalid-Plausible cases. In these instances, semantic truth occasionally overrides formal logic, causing the model to default to a 'Valid' prediction. Inspection of the generated reasoning traces reveals a pattern of "logical hallucination": to bypass structural constraints like the *Illicit Major* fallacy, the model actively fabricates distribution rules (e.g., falsely asserting that an I-type premise contains a distributed subject) and hallucinates classical valid forms (e.g., falsely labeling an invalid I-O-O structure with the valid name "Festino") to post-hoc rationalize the flawed deduction. A complete reasoning trace of this behavior is detailed in Appendix C.

Impact of Model Scale. Comparing the 8B and 14B architectures reveals critical capacity thresholds. First, the 14B baseline marginally outperforms the 8B baseline (36.62 vs. 36.09), indicating that without structured data, scaling parameters merely saturates shallow statistical mapping rather than enabling formal logic. Second, the 8B model collapses when fine-tuned on reasoning corpora (27.88). It lacks the parameter density to simultaneously process the rigid formatting of the four-step audit and the underlying boolean constraints. This confirms that mitigating belief bias requires a strict synergy of sufficient parameter capacity and targeted neuro-symbolic data.

6 Conclusion

We presented a teacher–student distillation framework for SemEval-2026 Task 11, combining DeepSeek-based logical auditing, CoT-augmented data generation, and LoRA fine-tuning of Qwen3-14B. Our approach achieves strong performance across both subtasks, reaching 96.86% accuracy and a ranking score of 39.81 on Subtask 1, and 26.02 on Subtask 3. Comprehensive error analysis shows that while content effect is substantially mitigated, addressing invalid-plausible predictions remains a central challenge. These results suggest that structured distillation and reasoning-oriented augmentation are effective but not sufficient for fully disentangling logical validity from plausibility bias.

Limitations

Despite relatively strong performance, our system has several limitations. First, plausibility-driven bias is reduced but not fully eliminated. Second, while DeepSeek-based auditing improves label reliability, it introduces dependency on external reasoning models whose biases may propagate to the student model. Third, due to computational constraints restricting our experiments to the 14B parameter scale, our investigation into how larger model capacities internalize complex CoT reasoning remains to be further explored.

References

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. *Preprint*, arXiv:2406.11341.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell,

Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models show human-like content effects on reasoning tasks. *Preprint*, arXiv:2207.07051.

J. St. B. T. Evans, Julie L. Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-R1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Geonhee Kim, Marco Valentino, and André Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. *Preprint*, arXiv:2408.08590.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. Exploring reasoning biases in large language models through syllogism: Insights from the neubaroco dataset. *Preprint*, arXiv:2408.04403.

Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. 2024. Verification and refinement of natural language explanations through llm-symbolic theorem proving. *Preprint*, arXiv:2405.01379.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 17222–17240. Association for Computational Linguistics.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2026a. Mitigating content effects on reasoning in language models through fine-grained activation steering. *Preprint*, arXiv:2505.12189.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026b. Semeval-2026 task 11: Disentangling content and formal reasoning in large language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. 2025. SylloBio-NLI: Evaluating large language models on biomedical syllogistic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7235–7258, Albuquerque, New Mexico. Association for Computational Linguistics.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *Preprint*, arXiv:2405.18357.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Appendix

A Hyperparameter Settings

Table 5: Hyperparameter settings for LoRA fine-tuning on Qwen-3 models. The settings were applied uniformly across all experiments.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-4}
Epochs	5
Early Stopping	3
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Target Modules	all-linear
Batch Size	4
Gradient Accumulation Steps	4

B Formal Auditing Pipeline and Reasoning Trace Example

Table 6 provides a concrete example of how the Auditor transforms natural language into a structured reasoning trace by (1) abstracting terms into symbolic variables, (2) identifying the syllogistic figure, (3) verifying the argument against classical logical rules, and (4) synthesizing a final judgment.

C Error Case Analysis: Logical Hallucination

Table 7 illustrates a severe case of "logical hallucination" from the evaluation set under the Invalid-

Table 6: Representative example of the distilled structured reasoning (CoT) trace from our training set.

Component	Content
Syllogism	*Every parrot has a beak. All parrots are classified as birds. Therefore, it must be that all birds have beaks.*
Distilled CoT	<think> **Symbolic Conversion:** - S: birds, M: parrots, P: things with beaks - Premise 1 Form: All M are P (A-type); Premise 2 Form: All M are S (A-type) **Step-by-Step Derivation:** Step 1: Figure 3 (M is Subject in both Premise 1 and Premise 2). Step 2: Distributed Middle Rule: M ("parrots") is distributed in both premises → Pass. Step 3: **Illicit Process Rule:** S ("birds") is distributed in Conclusion (A-type subject) but undistributed in Premise 2 (A-type predicate) → **Fail (Illicit Minor)**. Step 4: Negative Premises Rule: No negative premises → Pass. </think> **Validity Conclusion:** Invalid: Illicit process of the minor term (S distributed in conclusion but not in any premise).

Plausible condition. To justify an empirically plausible but formally invalid conclusion, the model fabricates boolean constraints and invents a non-existent valid syllogistic form to bypass the *Illicit Major* fallacy.

Table 7: A representative failure case demonstrating logical hallucination. The model's complete reasoning trace is provided, with fabricated constraints and hallucinated forms highlighted in bold.

Component	Content
Syllogism	*It is the case that some pencils are white. There are some sheets of paper that are not white. This implies that some sheets of paper are not pencils.*
Gold Label	**Invalid** (Fallacy of Two Particular Premises; Illicit Major)
Prediction	**Valid** (Incorrect)
Hallucinated CoT	<think> **Symbolic Conversion:** - S: sheets of paper - M: white - P: pencils - Premise 1 Form: Some P are M (I-type); Premise 2 Form: Some S are not M (O-type) **Step-by-Step Derivation:** Step 1: Figure 2 (M is Predicate in both premises). Step 2: Distributed Middle Rule: M ("white") is distributed in Premise 2 (O-type predicate) $\rightarrow$ Pass. Step 3: **Illicit Process Rule:** P ("pencils") is distributed in Conclusion (O-type predicate) and **distributed in Premise 1 (I-type subject)** $\rightarrow$ Pass. Step 4: Negative Premises Rule: One negative premise (Premise 2) and negative conclusion $\rightarrow$ Pass. </think> **Validity Conclusion: Valid: Follows the form I-O-O-2 (Festino)** with no syllogistic rule violations.