Thinking in Schemas: Robust Syllogistic Reasoning in LLMs

Federico Ranaldi[†,•] **Leonardo Ranaldi**[•]
Fabio Massimo Zanzotto[†,⋆] **Shay B. Cohen**[•]
[†] Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
[•] School of Informatics, University of Edinburgh, UK
[⋆] Almawave S.p.A., Italy
{first_name.last_name}@uniroma2.it

Abstract

LLMs often mistake what sounds true for what is formally valid. This limitation is especially evident in syllogistic reasoning, where plausible arguments can lead models to endorse conclusions that are logically invalid, a phenomenon known as content effect (CE).

We present *Boethius*, a schema-guided framework for syllogistic reasoning that disentangles semantic plausibility from logical validity. *Boethius* adopts an auditable, quasi-formal reasoning process with two complementary stages: a *Schema Module*, which deduces the underlying logical form by analysing the formal structure of the premises, and an *Instantiation Module*, which instantiates this form over the concrete argument and evaluates validity independently of content-level semantics.

Our results show that *Boethius* consistently outperforms existing approaches, improving syllogistic reasoning accuracy while substantially reducing CE. These gains hold for both large models in a pure in-context learning setting and smaller models trained via schema-guided trajectories using supervised fine-tuning and optimisation-based refinement.

1 Introduction

From Boethius to large language models (LLMs), the challenge of reasoning by form rather than belief has persisted. In De Syllogismo Categorico (ca. 520 CE), Boethius systematised Aristotelian syllogistic theory for the Latin world, explicitly separating logical necessity from persuasive plausibility. Fifteen centuries later, LLMs confront a strikingly similar problem: distinguishing what is formally valid from what merely sounds true (Eisape et al., 2024; Bertolazzi et al., 2024).

Improving the ability of LLMs to perform syllogistic reasoning[1] remains a key open problem for formal inference in natural language. Syllogisms provide a controlled environment where logical validity can be isolated from semantic plausibility and prior beliefs, making them a standard diagnostic tool for analysing reasoning failures in both humans and models (Eisape et al., 2024). This distinction has proven essential across applied domains, from medical and normative reasoning (Ozeki et al., 2025; Sim and Chen, 2025) to scientific discovery (Wysocka et al., 2025) and socio-political judgement (Keller et al., 2024; Aspernas et al., 2022; Aslanov et al., 2025).

Yet, despite their exposure to vast textual data, LLMs still conflate plausibility with validity, endorsing arguments that align with world knowledge or stored textual facts (Zanzotto et al., 2025) but violate formal rules of inference, suffering from a phenomenon known as content effect (Bertolazzi et al., 2024). Unlike typical natural language inference tasks, syllogistic inference demands alignment with abstract argument schemas, providing a precise probe of whether models are able to follow form-driven reasoning.

We fill this gap and present *Boethius*, a schema-guided framework for syllogistic reasoning. *Boethius* decomposes inference into two schema-centred stages: *(i) Schema Identification*, in which a **Schema Module** analyses the quantification and polarity structure of the premises to recover the underlying syllogistic form; and *(ii) Schema Instantiation*, in which a **Instantiation Module** applies the identified form to the argument and evaluates its validity. Together, these two stages enable a structured reasoning procedure that can be adopted at inference time.

Building on this formulation, we investigate whether such schema-guided structured reasoning can be transferred from large to small models. To this end, we collect complete dual-stage trajectories produced by the **Schema Module** and **Instantiation Module** and use them to supervise

[1] Here, we are referring to the general category of deductive inference where a small number of premises is given, and from that a logical conclusion is derived.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40891–40909

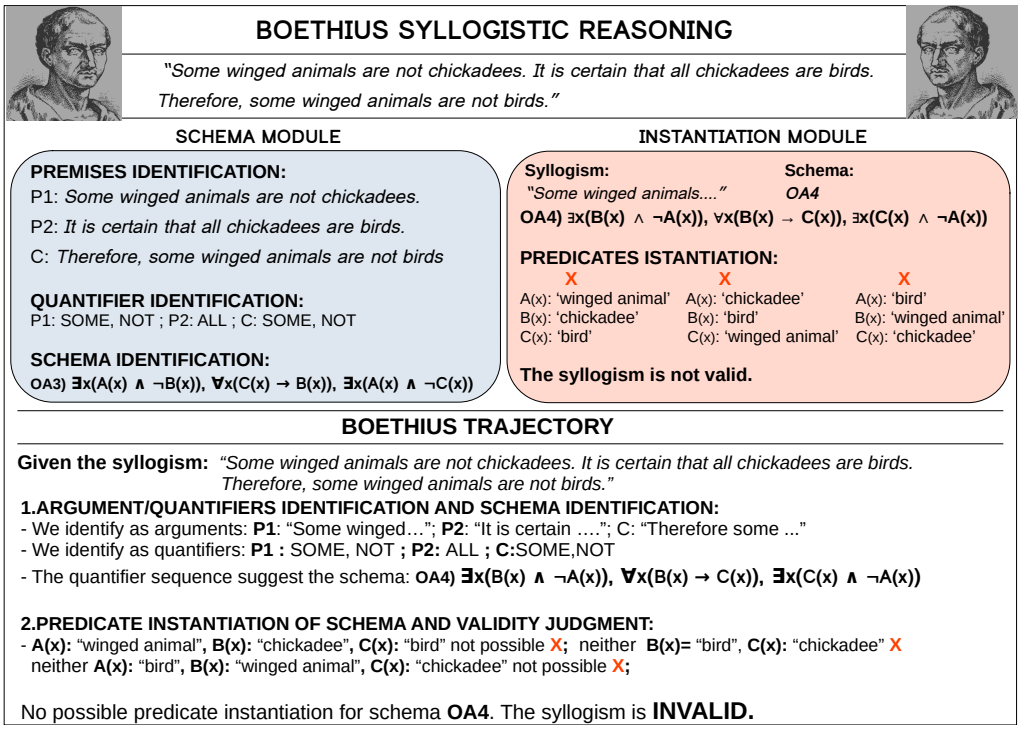

Figure 1: Illustration of the *Boethius* dual-stage reasoning process. The **Schema Module** identifies the underlying syllogistic form by operating on structural cues only, while the **Instantiation Module** applies the identified schema to determine whether the conclusion is valid or invalid.

smaller models through distillation, combining supervised fine-tuning with optimisation-based refinement. This setup enables small models to internalise content-independent, schema-guided reasoning strategies and eventually execute a full dual-stage pipeline, without requiring explicit external guidance.

Our contributions are threefold:

- **Boethius:** We introduce a dual-stage method for syllogistic reasoning that separates *Schema Identification* from *Schema instantiation*. By comparing our approach with deductive and quasi-symbolic approaches, we show that schema-guided reasoning leads to improved accuracy and a substantial reduction in content effect.
- **Distillation of structured reasoning:** Through supervised fine-tuning and optimisation on dual-stage trajectories, we show that smaller LLMs acquire more robust reasoning strategies, yielding compact models that implicitly plan and execute a dual-stage inference procedure.
- **Generalisation beyond syllogistic formalism:** *Boethius* preserves strong performance when evaluated on reasoning tasks that are not explicitly governed by syllogistic schemas, indicating that schema-guided supervision induces reasoning behaviours that generalise beyond the target formalism.

2 Method

We formulate syllogistic reasoning as a supervised classification task over natural language arguments, as done in the more general task of textual entailment recognition (Dagan et al., 2013; Pazienza et al., 2006). Each instance consists of two premises P_1 (*Some chickadees are not winged animals*), P_2 (*It is certain that all chickadees are birds*) and a conclusion C (*Therefore some birds are not winged animals*), all written in natural language. The model receives as input $x = \langle P_1, P_2, C \rangle$ and must predict a validity label $y \in \{\text{VALID}, \text{INVALID}\}$ indicating whether the conclusion follows from the premises according to syllogistic rules (Bertolazzi et al. 2024; see Appendix I). Many arguments are designed to induce a *content effect*: the conclusion is semantically plausible but formally invalid (or vice versa), so that purely plausibility-based heuristics are insufficient or misleading (Bertolazzi et al., 2024; Valentino et al., 2025; Maraia et al., 2026).

We claim that genuine syllogistic reasoning requires two skills: *(i)* identifying the underlying logical *schema* that structures the argument, and *(ii)*

evaluating the argument by applying this schema in a way that is independent of world knowledge and semantic plausibility.

2.1 Boethius Syllogistic Reasoning

To disentangle schema recognition from its application, we propose *Boethius*, a schema-guided syllogistic reasoning framework.

Given an input syllogism x, inference process is decomposed into:

1. *Schema Identification*: a **Schema Module** abstracts the logical structure of x and identifies the syllogistic schema that is most aligned with the argument.
2. *Schema Instantiation*: an **Instantiation Module** applies the identified schema to the concrete premises and produces the final validity judgement.

By centring both stages on schema, *Boethius* enforces a structured reasoning process in which validity judgements are derived from form rather than surface plausibility.

Schema Module is instructed to abstract away from content-specific plausibility and operate on the premises structure. Given x, it produces: an explanation of how the premises and the conclusion relate in terms of quantification structure; a syllogistic schema s encoding the corresponding quantifier pattern and figure (one of the logical schemas listed in Table 12).

To promote structural abstraction, the module is guided by an instruction template that requests the identification of key terms and their logical relations, and then maps this structure to a schema within a fixed inventory of syllogistic forms. The output is constrained to the format $\langle \text{explanation}, s \rangle$, where the explanation provides a concise natural-language justification and s is the schema label passed to the **Instantiation Module**.

Instantiation Module receives the input syllogism x together with the schema $\langle s \rangle$ produced by the **Schema Module**. Its role is to apply the rule encoded in s to the concrete argument in x and determine whether the conclusion is logically licensed by the premises.

The module outputs a pair $\langle \text{explanation}, y \rangle$, where $y \in \{\text{VALID}, \text{INVALID}\}$ denotes the resulting validity judgement. To minimise semantic leakage, the model is instructed to treat predicates as abstract placeholders and to prioritise structural conformity with the schema s over world-knowledge plausibility.

2.2 Trajectory Collection

To train our models, we collect reasoning trajectories in which each syllogism is paired with:

- a *Schema Identification* trajectory τ_{schema}, produced by the **Schema Module**;
- a *Schema Instantiation* trajectory τ_{inst}, produced by the **Instantiation Module**;
- a direct *Deductive* trajectory τ_{ded}, produced by a single-LLM Chain-of-Thought teacher.

Each trajectory combines the input syllogism, the intermediate reasoning steps, and the final validity judgement. In the *Boethius* setting, a complete reasoning trajectory is defined as:

$$\tau_{\text{dual}} = \langle x,\ \tau_{\text{schema}},\ s,\ \tau_{\text{inst}},\ y \rangle,$$

where τ_{schema} is the schema-level explanation generated by the **Schema Module**, s is the predicted syllogistic schema, τ_{inst} is the instantiation-based reasoning produced by the **Instantiation Module**, and $y \in \{\text{VALID}, \text{INVALID}\}$ is the final label.

The deductive trajectories τ_{ded} are instead used to train syllogistic reasoners following a traditional *Deductive* strategy, and serve as a comparison point for schema-guided training with *Boethius*.

We use these trajectories to train a single compact *syllogistic reasoner* that implicitly internalises the reasoning procedure and executes it in a single forward pass at inference time.

2.3 Distillation

We investigate whether syllogistic reasoning trajectories can be used to instruct smaller models via a combination of supervised fine-tuning (SFT) and preference optimisation.

Supervised Fine-Tuning As cold-start initialisation, we perform SFT on the collected trajectories, conditioning the student models to reproduce the teacher's structured outputs. We train the models for *Schema Identification* and *Schema Instantiation*, each with task-specific instruction templates. We concatenate the two stages into a unified, tagged format (e.g. `<schema>...</schema><instantiation>...</instantiation><answer>...</answer>`) and train the model to generate the full pipeline.

GRPO-based Refinement. Building on recent work on reasoning-oriented RL (Shao et al., 2024), we then refine the SFT models using group relative policy optimisation (GRPO), combining: Answer reward: correctness of the final validity label. Format reward: adherence to the required schema/instantiation structure. Abstraction reward: penalties for directly copying surface lexical content (to encourage symbol-like reasoning). Details in Appendix G.

3 Experiments

3.1 Dataset

We conduct all experiments on a controlled English corpus of syllogistic arguments. We build on the dataset proposed by Bertolazzi et al. (2024), which follows the standard taxonomy of categorical syllogistic forms adopted in the cognitive reasoning literature (detailed in Appendix I).

Each schema is expressed in a textual first-order logic formalism. To control lexical content, templates are instantiated using noun phrases sampled from WordNet taxonomic relations, thereby ensuring a consistent hypernym–hyponym structure. Each instance is represented as a string $x = \langle P_1, P_2, C \rangle$ and is annotated with two attributes: INVALID/VALID, indicating whether the conclusion follows under the target schema; and UNBELIEVABLE/BELIEVABLE, indicating whether the conclusion accords with world knowledge. These two dimensions enable the systematic construction of believable–plausible and implausible–invalid pairs, which are central to probing content effects. Examples and details are reported in Appendix A.

The resulting corpus comprises 10,000 balanced arguments, covering all configurations. We split the set, using 40% for fitting policies and a **held-out evaluation set** for reporting results. Unlike Valentino et al. (2025), we avoid explicit "Premise 1/2" markers, as they could introduce positional shortcuts. Presenting the argument as a single string requires the model to reconstruct its structure, which aligns with our schema-reconstruction objective.

3.2 Metrics

We evaluate all models on a held-out test set (§3.1), comparing our method *Boethius* against *Baseline* prompting, a *Deductive* strategy, and a *Quasi-Symbolic* approach (Quasar). These comparative approaches are outlined below in §3.3 and in Appendix H. Our metrics quantify both overall reasoning performance and the extent to which models resist content-driven heuristics.

Accuracy (Acc). We compute standard accuracy over the VALID/INVALID labels. Accuracy is reported both *globally* and *per Believability*.

Content Effect (CE). To quantify the extent to which semantic plausibility interferes with logical validity judgements, we measure the CE by focusing on the two cases: *Invalid–Believable* and *Valid–Unbelievable*. Specifically, CE is defined as: $\big((1 - \text{Invalid}_{Bel}) + (1 - \text{Valid}_{Unbel})\big)/2$, where Invalid_{Bel} denotes accuracy on Invalid–Believable arguments, and Valid_{Unbel} accuracy on Valid–Unbelievable arguments. Lower CE values indicate stronger resistance to content-driven biases, reflecting a model's ability to prioritise logical validity over semantic plausibility.

Schema-level Metrics. Since our approach explicitly separates *Schema Identification* from *Schema Instantiation*, we report: **Schema Identification Accuracy**: number of instances for which the **Schema Module** predicts the correct logical schema. **Schema Fidelity**: number of **Instantiation Module** predictions that are consistent with the label implied by the schema predicted in the first stage. This metric evaluates whether the two modules operate coherently.

Dual-stage vs. Single-stage Ablations. For the distilled models, we also report the proportion of generated reasoning traces whose internal structure matches the expected dual-stage format (§ 2.3). Faithful trace structure is evaluated via heuristic defined in Appendix C and reported as **Structure Fidelity**: proportion of traces in which *Schema Identification* and *Schema Instantiation* are both valid. High **Structure Fidelity** indicates that small models have internalised the two-stage reasoning procedure.

3.3 Deductive and Quasi-Symbolic Pipelines

As comparison points, we consider three strategies capturing increasing levels of reasoning guidance: a *Baseline* with no explicit guidance, a *Deductive* strategy to test whether free CoT reasoning improves validity judgements, and a *Quasi-Symbolic* strategy (Ranaldi et al., 2025) that steers the model toward structured abstractions without enforcing full modular decomposition.

Model	Believable		Unbelievable		CE
	All	Invalid	All	Valid	
Llama3-1B	60.8	38.5	46.8	9.3	76.1
+deductive	63.5	40.4	48.6	11.4	74.1
+quasar	66.1	40.2	49.3	12.6	73.6
+Boethius$_{Decomposed}$	**72.1**	**48.4**	**56.0**	**21.8**	**64.9**
Llama3-8B	66.1	42.0	51.0	11.5	73.2
+deductive	68.6	46.0	55.4	19.1	67.5
+quasar	69.6	47.2	57.2	22.0	65.4
+Boethius$_{Decomposed}$	**80.2**	**63.0**	**66.4**	**35.6**	**50.7**
Llama3-70B	71.8	51.2	57.9	21.8	63.5
+deductive	73.0	52.2	60.6	25.4	61.2
+quasar	75.4	55.8	60.3	27.2	58.5
+Boethius$_{Unified}$	**82.1**	**66.0**	***66.8***	**35.4**	**49.3**
Gemma2-9B	67.2	50.4	57.9	24.1	62.8
+deductive	69.8	51.6	60.3	26.4	61.4
+quasar	**70.5**	52.0	60.8	27.0	60.5
+Boethius$_{Decomposed}$	72.6	**54.8**	**62.8**	**31.8**	**56.7**
Qwen-3-1B	58.2	37.5	46.1	8.9	76.8
+deductive	61.9	39.5	48.2	11.0	74.8
+quasar	61.2	40.5	51.9	18.0	70.8
+Boethius$_{Decomposed}$	**71.3**	**46.6**	**60.6**	**28.2**	**62.6**
Qwen3-8B	68.4	50.2	57.0	23.3	63.2
+deductive	66.2	44.2	59.3	24.6	65.6
+quasar	71.0	50.0	60.6	26.8	61.6
+Boethius$_{Decomposed}$	**71.6**	**53.1**	**62.2**	**28.9**	**59.0**
Qwen3-32B	71.2	51.0	57.3	21.4	63.8
+deductive	72.8	51.8	60.3	25.2	61.5
+quasar	73.2	52.3	61.1	26.9	60.4
+Boethius$_{Unified}$	***82.0***	**65.7**	**66.2**	**35.5**	**49.4**
GPT-4o	73.4	53.3	58.9	23.1	61.8
+deductive	74.0	53.9	61.7	27.7	59.2
+quasar	76.9	55.0	62.3	28.0	58.4
+Boethius$_{Unified}$	**81.9**	**66.2**	**65.7**	**33.4**	***50.2***

Table 1: Accuracy levels on *Invalid* and *Valid*, both (*All*) and Content Effect (*'CE'*) using methods introduced in §2. In bold, the best results per model.

3.4 Models

We evaluate our approach on a range of open and proprietary LLMs, spanning different scales and training regimes: Llama3 (1B, 8B, 70B; Grattafiori et al. 2024), Gemma 2-9B (Team et al., 2024), Qwen 3 (1B, 7B, 72B; Yang et al. 2025) and GPT-4o (OpenAI et al., 2024).

4 Results and Discussion

We evaluate *Boethius* with respect to in-context learning (ICL) accuracy (§4.1), robustness to content effect (§4.2), and trajectory distillation for small models §4.3, analysing both accuracy and CE metrics.

4.1 Reasoning Strategies

As shown in Table 1, moving from unstructured to increasingly structured and guided strategies yields substantial gains in syllogistic reasoning accuracy across all models. *Baseline* performs weakest overall, with Believable Accuracy ranging between 58–73% and consistently poor performance on Unbelievable cases (e.g., 46.8% for Llama3-1B and 46.1% for Qwen-3-1B). *Deductive* strategy provides non-uniform accuracy changes across models, showing modest gains on larger models while slightly degrading performance on smaller ones. The quasi-symbolic approach, *Quasar*, further improves accuracy by enforcing a structured reasoning pipeline, with pronounced gains on Unbelievable cases.

Since smaller models exhibited hallucinations when outputs exceeded a certain length, we apply the *Boethius* strategy in a decomposed form, splitting the two reasoning stages into separate prompts executed within the same context. We refer to these variants as *Boethius*$_{Unified}$ and *Boethius*$_{Decomposed}$.

Across all model families, the highest accuracy levels are achieved by *Boethius*. For smaller models, *Boethius*$_{Decomposed}$ yields significant absolute gains over both *Deductive* and the *Quasar*. Llama3-1B improves from 66.1% (*Quasar*) to 72.1% in Believable cases, and Llama3-8B shows a similar trend that is also more pronounced in Unbelievable cases (57.2% (*Quasar*) to 66.4% (*Boethius*)). For larger models, Boethius$_{Unified}$ consistently achieves the best performance, reaching more than 81% Believable accuracy for Llama3-70B, Qwen3-32B, and GPT-4o. Overall, these results indicate that guiding models through explicit schema-centred stages provides a stronger inductive signal than purely *Deductive* or quasi-symbolic approaches, and that the most effective *Boethius* configuration depends on model scale.

4.2 Robustness to the Content Effect

Table 1 (last column) shows clear differences across strategies in terms of content effect. *Baseline* prompting exhibits a consistent CE across all models, with values often exceeding 70% for 1B models, reflecting a heavy reliance on semantic plausibility rather than logical form.

The *Deductive* strategy moderately reduces CE, with the exception of Qwen-3-7B, where CE slightly increases. *Quasar*, further mitigates CE, yielding more consistent reductions across model

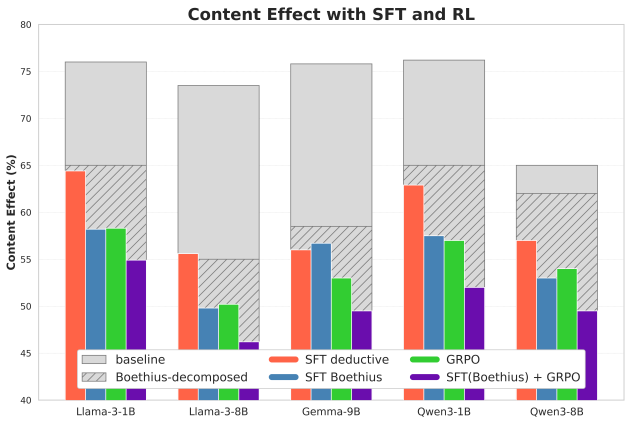

Figure 2: Content Effect scores using baseline and different tuning methods reported in §2.3.

sizes (e.g., for Llama3-8B, CE decreases from 73.2% to 65.4%), suggesting that quasi-symbolic abstraction partially attenuates content-driven bias.

The most substantial CE reductions are achieved by *Boethius*. For smaller models, $Boethius_{Decomposed}$ reduces significantly CE with respect to *Quasar* (e.g., from 73.6% to 64.9% for Llama3-1B and from 70.8% to 62.6% for Qwen-3-1B). For 8B-scale models such as Llama3-8B, *Boethius* yields a markedly larger CE reduction than *Quasar* (65.4% to 50.7%). For larger models, $Boethius_{Unified}$ achieves the lowest CE across all models, ranging from approximately 49–50% on Llama3-70B, Qwen3-32B, and GPT-4o.

Generally, these results show that *Boethius* consistently suppresses CE, shifting validity judgements toward schema-guided reasoning. While gains for small models are decisive, they remain bounded (especially for 1B models); therefore, in §4.3 we examine trajectory distillation and the role of training and optimisation strategies.

4.3 Trajectory Distillations

Figure 2 and Table 10 in Appendix F show that distilling trajectories (generated from GPT-4o) substantially improves the performance of small models, yielding consistent gains over both *Baseline* and *Deductive* strategies. Among distillation approaches, SFT on *Deductive* trajectories performs worst. Notably, for Llama-3-8B, *Boethius* prompting already exhibits lower content effect than SFT(*Deductive*). Table 10 further shows numerically that CE is consistently highest under the *Baseline* setting and decreases systematically when *Boethius* schema-guided distillation is applied. The lowest CE values are obtained by combining SFT and GRPO, as defined in §2.3. Accuracy levels in Table 10 confirm this trend across all small models (Llama-3-1B/8B, Gemma2-9B, Qwen-3-1B/8B): distillation on *Boethius* trajectories consistently outperforms distillation on *Deductive* trajectories for both Believable and Unbelievable syllogisms. For instance, Llama-3-1B reduces CE from 63.9% under SFT(*Deductive*) to 58.3% under SFT(*Boethius*), while Unbelievable accuracy increases from 53.9% to 62.3%.

Regarding training methods, the combination of SFT and GRPO delivers the strongest overall results. As shown in Figure 3, this is the only setting in which both reward score and accuracy increase steadily with training steps, whereas alternative methods exhibit early saturation or overfitting. This configuration achieves the lowest CE across all small models (looking at Table 10, 54.5% for Llama-3-1B, 48.8% for Llama-3-8B, and 53.5% for Qwen-3-1B), while also yielding the highest accuracy on Unbelievable syllogisms. Overall, these findings indicate that *Boethius* trajectories provide a substantially stronger supervision signal than free-form *Deductive*, with SFT+GRPO proving especially effective for transferring structured syllogistic reasoning to small models.

4.4 Component-level Analysis

To better understand the sources of residual errors and content effect, we report a set of diagnostic analyses in Appendix C. We evaluate **Schema Identification Accuracy** independently of final validity judgements and analyse the internal structure of reasoning traces. Across models, **Schema Identification Accuracy** is consistently high, indicating that most remaining errors arise during *Schema Instantiation*, where semantic interference plays a key role. Trace-level analyses in Appendix 8 further show that *Boethius* distillation improves both **Schema Fidelity** and **Structural Fidelity**, with distilled models more reliably reproducing the intended dual-stage format. Additionally we conduct a answer-level diagnostics (reported in Appendix E), showing that schema-guided distillation substantially improves the consistency of final validity outputs.

4.5 Qualitative Error Analysis

We provide qualitative analysis on examples of Unbelievable–Valid syllogisms for GPT-4o and Qwen-3-32B (Appendices J–K).

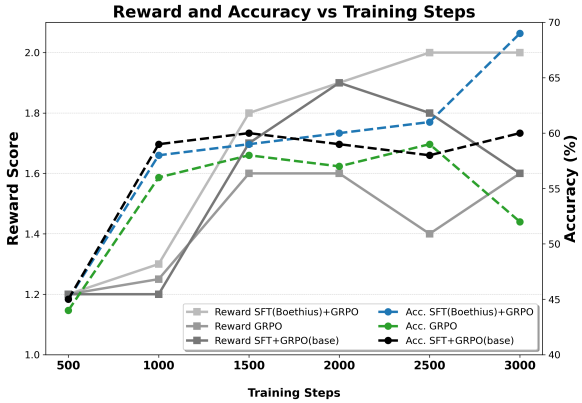

Figure 3: Study on Reward Score and average Accuracy on Llama-3-8B.

Model	Avg	En	Zh	De
Llama-3-8B	76.4	73.8	77.4	78.1
+Boethius$_{Decomp}$	54.1	49.9	57.1	55.3
SFT+GRPO	**45.8**	**45.8**	**46.0**	**45.6**
Qwen-3-8B	66.2	63.8	68.7	66.2
+Boethius$_{Decomp}$	59.0	57.8	60.5	58.7
SFT+GRPO	**50.1**	**49.9**	**50.6**	**49.8**

Table 2: Multilingual Content Effect (CE; ↓ lower is better). We report the average CE for English (En), Chinese (Zh), and German (De). *Boethius* distillation, completed with the combination of SFT and GRPO, substantially reduces content-driven biases across all languages.

Baseline prompting exhibits immediate content-driven errors: arguments are rejected solely because one or more premises contradict world knowledge, without any attempt to assess logical form. *Deductive* prompting amplifies this behaviour by explicitly justifying the rejection in terms of semantic coherence, demonstrating that free-form CoT does not prevent CE.

Applying *Boethius* in a single-stage configuration (*Boethius*$_{Unified}$) partially mitigates this issue. In both examples, the model correctly recovers the underlying schema (IE1). However, errors persist at the instantiation stage: for conclusions of the form *"Some roses are not flowers"* or *"No mammal is an animal"*, models allow taxonomic world knowledge to leak into the reasoning process, causing the rejection over otherwise valid instantiation.

Only *Boethius*$_{Decomposed}$ pipeline consistently yields correct judgements. By explicitly separating *Schema Identification* from *Schema Instantiation*, the model treats predicates as abstract placeholders during inference, preventing plausibility from contaminating validity. Notably, we observe that errors involving incorrect *Schema Identification* followed by correct *Schema Instantiation* are rare, whereas the inverse pattern is common in single-stage settings. This asymmetry supports our quantitative findings that schema-level abstraction is the primary driver of robustness to CE.

4.6 Ablations

Multilingual Performance Table 2 shows that, across English, Chinese, and German, the *Baseline* displays consistently high content effect for 8B models. Applying *Boethius*$_{Decomp}$ already yields a clear reduction in CE for both models, indicating that schema-guided prompting mitigates content sensitivity beyond English and transfers to typologically distinct languages.

The strongest reductions are obtained when *Boethius* trajectories are combined with SFT and GRPO. Under this setting, CE decreases consistently across all languages and models, yielding the lowest average scores and substantially narrowing cross-lingual variation. Notably, improvements are comparable across languages, suggesting that the benefits of schema-level abstraction are not language-specific but operate at a form-driven level. These results indicate that schema-guided distillation promotes language-agnostic resistance to CE, enabling more robust syllogistic reasoning across multilingual settings. Additional results on other languages are reported in §5.

Boethius Impact on Other Reasoning Tasks. Table 3 evaluates whether models trained on *Boethius* trajectories retain their reasoning capabilities on tasks beyond syllogistic inference. We report results on MMLU-Redux (Gema et al., 2025) and GSM-Symbolic (Mirzadeh et al., 2025), including perturbed variants designed to test robustness to surface-level changes. Across all evaluated models, training with *Boethius* does not degrade performance relative to baseline or CoT prompting, indicating that *Boethius* supervision does not interfere with general reasoning abilities.

On MMLU-Redux, performance under *Boethius* remains comparable to standard CoT approach across both original and choice-shuffled settings, with only marginal differences across models. This suggests that enforcing schema-level abstraction does not impair multi-choice reasoning or sensitivity to answer ordering. On GSM-Symbolic, *Boethius*-trained models consistently match or out-

perform both baseline and CoT, including under second-choice perturbations, indicating improved robustness in symbolic arithmetic reasoning.

Overall, these results demonstrate that *Boethius* training does not introduce task-specific overfitting. Instead, *Boethius* preserves, and in some cases enhances performance on diverse reasoning benchmarks, supporting the claim that its inductive bias is compatible with general reasoning rather than narrowly specialised to syllogistic inference.

	Task	Baseline	CoT	Our
Llama3-8B	MMLU-Redux	67.2	70.2	67.8
	-choices shuffled	63.0(**-3.2**)	68.0(**-2.2**)	67.6 (-0.2)
	GSM-Symbolic	49.5	**55.4**	50.2
	-2nd choice	44.0(**-4.5**)	48.4(**-1.8**)	50.0 (-0.2)
Qwen3-8B	MMLU-Redux	78.0	80.6	79.1
	-choices shuffled	76.0(**-2.0**)	78.2(**-1.2**)	78.9 (-0.2)
	GSM-Symbolic	46.7	60.2	65.3
	-2nd choice	44.9(**-1.8**)	58.4(**-1.8**)	64.8 (-0.5)
Gemma2-9B	MMLU-Redux	72.4	74.6	74.0
	-choices shuffled	70.2(**-1.8**)	74.0(**-0.6**)	73.3 (-0.7)
	GSM-Symbolic	48.9	62.6	67.0
	-2nd choice	48.0(**-0.9**)	62.4(**-0.2**)	67.0 (-0.0)

Table 3: Performance on original and perturbed MMLU-Redux and GSM-Symbolic. *(differences in brackets)

5 Background and Related Work

A growing body of work investigates methods for improving LLM performance on formal reasoning tasks, including mathematical, symbolic and logical inference.

LLMs are Soft-formal Reasoners Despite their strong linguistic competence, LLMs often exhibit logical inconsistencies, hallucinated derivations, and structurally invalid inferences when faced with abstract or formal reasoning tasks (Saparov and He, 2023; Dasgupta et al., 2024). Recent studies further show that, state-of-the-art models display human-like reasoning biases, such as belief bias and content effect, especially in tasks like syllogistic inference (Eisape et al., 2024). Complementary analyses further show that LLMs often conflate logical validity with semantic plausibility, revealing a persistent entanglement between formal and world knowledge (Bertolazzi et al., 2024).

Improving Formal and Symbolic Reasoning in LLMs Deductive strategies have been used to elicit slightly reliable reasoning paths with a small improvement in mathematical and symbolic domains (Jiang et al., 2025; Liu et al., 2023). Still, even with explicit reasoning steps, models continue to display human-like biases, most notably CE (Eisape et al., 2024; Bertolazzi et al., 2024).

In-context learning (ICL) offers a complementary approach: few-shot prompting help models better discriminate valid from invalid arguments. In syllogistic reasoning, Bertolazzi et al. (2024) show that both ICL and SFT partially mitigate CE by exposing models to diverse logical forms.

Other approaches reduce logical inconsistencies through hybrid pipelines, where models formalise natural-language problems into symbolic forms, delegate inference to external solvers, and translate the solution back into natural language (Jiang et al., 2024; Pan et al., 2023). Despite their effectiveness in controlled settings, these methods are constrained by challenging formalisation, lengthy proofs, and limited applicability to naturally posed problems.

These challenges have motivated lighter, pseudo-symbolic approaches such as Quasar (Ranaldi et al., 2025) and Aristotle (Xu et al., 2025), which guide models to adopt structured formalisms during the reasoning process. This design retains much of the stability of symbolic reasoning while remaining flexible for natural-language inputs.

Transferring Reasoning to Small Models. Early work on reasoning transfer focused on short, human-like CoT, while more recent studies have explored longer and more explicit reasoning trajectories that expose intermediate inferential steps (Wei et al., 2023; Longpre et al., 2023). To transfer these capabilities to smaller models, supervised fine-tuning on teacher-generated reasoning traces has proven effective, allowing compact models to internalise structured reasoning behaviours (Ranaldi and Freitas, 2024; Fu et al., 2023). More recently, reinforcement learning with structured rewards, such as GRPO and related policy-based frameworks, have been used to refine reasoning behaviours by encouraging coherence, rule consistency, and resistance to shortcut heuristics (DeepSeek-AI et al., 2025).

6 Conclusion

We introduced *Boethius*, a schema-guided framework for syllogistic reasoning, and showed that explicitly separating *Schema Identification* from *Schema Instantiation* yields consistent improvements in both accuracy and robustness to content effect. These findings indicate that LLMs benefit from inference processes guided by explicit,

schema-centred stages rather than undifferentiated end-to-end reasoning.

We use *Boethius* trajectories to distil small syllogistic reasoners that autonomously reproduce the two-stage reasoning process. Students trained on *Boethius* (via SFT, GRPO, or their combination) exhibit markedly reduced sensitivity to CE and more stable formal reasoning than those distilled from *Deductive* approach.

Overall, this work highlights formalisation-first reasoning as a general and effective approach for improving the logical robustness of LLMs, enabling reliable reasoning even in semantically challenging settings.

Limitations

This work focuses on syllogistic reasoning as a controlled setting for studying form-driven inference. While syllogisms are widely used in domains such as biomedical, legal, and socio-political reasoning, the proposed framework is not extensively evaluated on other forms of logical reasoning. Although our ablation analyses on additional reasoning tasks indicate that *Boethius* does not degrade, and in some cases improves, general reasoning performance, its full generalisability beyond the syllogistic setting remains to be systematically explored.

Multilingual evaluation is limited to a small number of high-resource, typologically diverse languages. Extending the analysis to low-resource languages would further strengthen conclusions about cross-lingual generalisability.

For invalid arguments, the method adopts a heuristic strategy by testing instantiation against the structurally closest schema rather than exhaustively refuting all valid schemas. This choice is sufficient for classification but does not constitute a full logical proof of non-validity.

Finally, our evaluation centres on model behaviour; a systematic comparison with human performance, particularly across *Schema Identification* and *Schema Instantiation* subtasks, could further clarify the cognitive alignment and limitations of schema-guided reasoning.

Acknowledgements

We thank Adi Simhi for serving as a pre-submission reviewer and for the helpful feedback provided on an earlier draft of this paper. We also thank the anonymous reviewers for their constructive comments, which helped improve the clarity and overall quality of the manuscript. We are additionally grateful to the colleagues and institutional representatives who supported and made possible the collaboration between the University of Rome Tor Vergata and the University of Edinburgh. This research has been partially funded by MeMo – Project no. FISA-2024-00249 under the Fondo Italiano per le Scienze Applicate (FISA).

References

Ivan Aslanov, Anita Tobar-Henríquez, and Ernesto Guerra. 2025. Prior beliefs impair logical reasoning about syllogisms on sexual violence.

Julia Aspernas, Arvid Erlandsson, Ernesto Guerra, and Artur Nilsson. 2022. Motivated formal reasoning: Ideological belief bias in syllogistic reasoning across diverse political issues.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13882–13905, Miami, Florida, USA. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen,

Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444, Mexico City, Mexico. Association for Computational Linguistics.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *Preprint*, arXiv:2301.12726.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? 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 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dongwei Jiang, Marcio Fonseca, and Shay Cohen. 2024. LeanReasoner: Boosting complex logical reasoning with lean. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7497–7510, Mexico City, Mexico. Association for Computational Linguistics.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2025. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *Preprint*, arXiv:2405.06705.

Lucas Keller, Felix Hazelaar, Peter M. Gollwitzer, and Gabriele Oettingen. 2024. Political ideology and environmentalism impair logical reasoning.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *Preprint*, arXiv:2304.03439.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. 2026. Can activation steering generalize across languages? a study on syllogistic reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2739–2753, Rabat, Morocco. Association for Computational Linguistics.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *Preprint*, arXiv:2410.05229.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2025. Normative reasoning in large language models: A comparative benchmark from logical and modal perspectives. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 276–294, Suzhou, China. Association for Computational Linguistics.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *Preprint*, arXiv:2305.12295.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2006. Learning textual entailment on a distance feature space. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3944 LNAI:240 – 260. Cited by: 22.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

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)*, pages 17222–17240, Vienna, Austria. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *Preprint*, arXiv:2210.01240.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Shamus Zi Yang Sim and Tyrone Chen. 2025. Critique of impure reason: Unveiling the reasoning behaviour of medical large language models.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *Preprint*, arXiv:2505.12189.

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, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. 2025. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3052–3075, Vienna, Austria. Association for Computational Linguistics.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Fabio Massimo Zanzotto, Elena Sofia Ruzzetti, Giancarlo A. Xompero, Leonardo Ranaldi, Davide Venditti, Federico Ranaldi, Cristina Giannone, Andrea Favalli, and Raniero Romagnoli. 2025. Position paper: MeMo: Towards language models with associative memory mechanisms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15169–15180, Vienna, Austria. Association for Computational Linguistics.

A Dataset

Syllogisms are split into 4 categories: Valid–Believable,Valid–Unbelievable,Invalid–Believable and Invalid–Unbelievable.

A.1 Valid and Invalid

Valid syllogisms are instantiated from the complete set of 27 classically valid categorical syllogistic schemas by assigning subject, middle, and predicate terms in accordance with each schema's inferential constraints, thereby yielding conclusions that are logically entailed by the premises. For instance, the following syllogism instantiates schema AA2 and is labelled as *Believable–Valid*: `'schema': 'AA2', 'x': 'All felines are cats. It is certain that all Siameses are felines. Therefore all Siameses are cats'.`

Invalid syllogisms are derived from their valid counterparts via *structure-preserving transformations*. In these cases, the quantifier pattern and lexical material are kept fixed, while the positional assignment of arguments is altered. This produces a syllogism that closely matches the valid instance in surface form but is no longer logically licensed by the premises, and is therefore labelled *Believable–Invalid*. For example, using the same lexical content: `'schema': 'AA2', 'x': 'All felines are cats. It is certain that all Siameses are cats. Therefore all Siameses are felines'.`

A.2 Believable and Unbelievable

Believability is manipulated independently of logical validity. Believable syllogisms are constructed by instantiating otherwise identical schemas with noun phrases drawn from coherent taxonomic relations in WordNet, such as hypernym–hyponym pairs (e.g., "All Siameses are felines") or closely related lexical variants. This results in premises and conclusions that are consistent with common world knowledge.

Unbelievable syllogisms, by contrast, are generated by introducing taxonomic inversions or semantically incompatible term assignments, while preserving grammatical well-formedness and the underlying schematic structure. As a result, believability varies independently of both schema and validity. For example, the following syllogism instantiates schema AA2 and is labelled as *Unbelievable–Valid*: `'schema': 'AA2', 'x': 'All cats are dogs. It is certain that all felines are cats. Therefore all felines are dogs'.`

Conversely, an *Unbelievable–Invalid* instance can be obtained by altering the positional assignment of terms while retaining the same lexical material: `'schema': 'AA2', 'x': 'All cats are dogs. It is certain that all felines are dogs. Therefore all felines are cats'.`

A.3 Multilingual Syllogistic Data

Starting from the original version in English (en), we construct multilingual versions in German (de), Spanish (es), French (fr), Italian (it), and Chinese (zh). We translate into the target language using GPT-4 and then apply back-translation to English.

Translation Quality Check We perform a three-stage quality control to preserve both *form* (logical structure, quantifiers, polarity) and *content* (lexical semantics, named categories):

Structural and lexical guards We enforce template-level invariants on determiners and quantifiers (*all/no/some/some... not*), prohibit polarity drift, and check lemma preservation for key nouns via bilingual lexicons. We require high semantic similarity between source and back-translation (SBERT cosine ≥ 0.90) and surface consistency (chrF ≥ 0.70).

Bilingual chech Native/near-native reviewers verify *(i)* faithful rendering of quantifiers, *(ii)* absence of idiomatic changes that could alter plausibility, and *(iii)* strict isomorphism of logical form. Disagreements trigger a corrective re-prompt with targeted constraints.

Logical invariance check We validate that *validity* and *plausibility* labels remain unchanged after translation/back-translation.

Language	SBERT Cosine	chrF	Pass Rate (%)
English (en)	0.94	0.81	100.0
German (de)	0.92	0.76	97.8
Spanish (es)	0.93	0.77	98.2
French (fr)	0.91	0.74	96.5
Italian (it)	0.93	0.79	97.1
Chinese (zh)	0.89	0.71	94.8

Table 4: Translation quality and pass rates per language.

Model	Avg	It	Fr	Es
Llama-3-8B	77.7	76.0	77.8	79.5
+Boethius$_{Decomp}$	54.8	50.2	57.8	56.6
SFT+GRPO	**45.8**	**45.2**	**46.4**	**46.0**

Table 5: CE for Italian (It), French (Fr), and Spanish (Es). *Boethius* completed with SFT and GRPO.

B Models & Parameters

We propose different models (detailed in Table 6). We choose the generation temperature for (mostly) deterministic outputs, with a maximum token length related to our strategy. The other parameters are left unchanged as recommended by the official resources. We use four 48GB NVIDIA RTXA600 GPUs for all experiments. In §2, we described the standard RL setting.

Model	Version
Gemma-2-9B	google/gemma-2-9b
Llama-3-1/8/70B	meta-llama/Meta-Llama-3-1/8/70B
Qwen3-1/8/32B	Qwen/Qwen3-1/8/32B
GPT-4o	gpt-4o-2024-08-06
GPT-4o-mini	gpt-4o-mini-2024-07-18

Table 6: Models used, which can be found on huggingface.co. We used all the default configurations proposed in the repositories for each model.

C Schema Identification Accuracy

Since our framework decomposes syllogistic reasoning into *Schema Identification* and *Schema Instantiation*, we evaluate the models' ability to recognise the underlying logical schema independently of the final validity judgement. We compute the *Schema Identification Accuracy* by comparing the schema label predicted in the `<schema>` block with the ground-truth syllogistic schema associated with each instance (defined by figure), independently of whether the final validity judgement is correct. Predictions are counted as correct if the identified schema exactly matches the gold schema, regardless of the correctness of subsequent instantiations or conclusions. Table 7 shows consistently high accuracy, indicating that both large and smaller LLMs retrieve the correct syllogistic form from the quantifier and polarity structure of the argument. These results demonstrate that residual errors and CE do not arise from failures in *Schema Identification* phase but from problems in instantiating the schema or in overriding semantic plausibility.

Model	no-tuning	SFT+GRPO
Llama-3-1B	84.9	94.2 (+9.3)
Llama-3-8B	94.6	96.0 (+8.5)
Llama-3-70B	98.1	-
Gemma-2-9B	95.9	96.8 (+0.9)
Qwen-3-1B	85.8	90.6 (+4.8)
Qwen-3-8B	96.2	97.4 (+1.2)
Qwen-3-32B	97.9	-
GPT-4o	98.4	-

Table 7: Schema recognition accuracy across models

D Qualitative Trace-level Metrics

To complement accuracy and CE (§4), we report trace-level metrics that describe the internal structure and qualitative properties of the generated trajectories. **Schema Accuracy (SchAcc)** measures correct schema identification, **Schema Fidelity (SchFid)** captures consistency between predicted schema and instantiation, and **Structure Fidelity (StrFid)** assesses whether traces follow the expected dual-stage format.

Model	SchAcc	SchFid	StrFid
Llama-3-1B$_{Unified}$	82.0	78.0	-
Llama-3-1B$_{Decomposed}$	84.9	82.1	-
Llama-3-1B$_{SFT+GRPO}$	94.2	86.4	84.2
Llama-3-8B$_{Unified}$	94.6	86.0	-
Llama-3-8B$_{Decomposed}$	94.6	93.2	-
Llama-3-8B$_{SFT+GRPO}$	96.0	98.0	90.7
Gemma-2-9B$_{Unified}$	95.9	88.4	-
Gemma-2-9B$_{Decomposed}$	95.9	92.0	-
Gemma-2-9B$_{SFT+GRPO}$	96.8	96.8	96.8
Qwen-3-1B$_{Unified}$	85.8	77.5	-
Qwen-3-1B$_{Decomposed}$	85.8	88.4	-
Qwen-3-1B$_{SFT+GRPO}$	90.6	92.0	87.2
Qwen-3-8B$_{Unified}$	96.2	87.6	-
Qwen-3-8B$_{Decomposed}$	96.2	94.8	-
Qwen-3-8B$_{SFT+GRPO}$	97.4	98.0	96.0

Table 8: Schema-/trace-level qualitative metrics.

E Answer Explicitness and Uniqueness

Answer Explicitness (AnsExp) measures the proportion of traces that contain an explicit final validity label (`VALID` or `INVALID`), recoverable via exact string matching. **Answer Uniqueness (AnsUniq)** measures the proportion of traces in which exactly one such validity label occurs.

Model	AnsExp	AnsUniq
Llama-3-1B$_{Unified}$	71.4	65.2
Llama-3-1B$_{Decomposed}$	78.9	73.8
Llama-3-1B$_{SFT+GRPO}$	96.1	93.5
Llama-3-8B$_{Unified}$	82.7	78.9
Llama-3-8B$_{Decomposed}$	89.5	86.7
Llama-3-8B$_{SFT+GRPO}$	98.4	97.2
Gemma-2-9B$_{Unified}$	85.2	81.4
Gemma-2-9B$_{Decomposed}$	91.8	89.6
Gemma-2-9B$_{SFT+GRPO}$	99.1	98.6
Qwen-3-1B$_{Unified}$	74.6	69.1
Qwen-3-1B$_{Decomposed}$	86.3	83.0
Qwen-3-1B$_{SFT+GRPO}$	97.0	95.4
Qwen-3-8B$_{Unified}$	88.9	85.7
Qwen-3-8B$_{Decomposed}$	93.7	91.9
Qwen-3-8B$_{SFT+GRPO}$	99.0	98.2

Table 9: Answer Explicitness and Answer Uniqueness.

F Detailed Results

Model	Believable			Unbelievable			CE
	Acc	Valid	Invalid	Acc	Valid	Invalid	
Llama3-1B	60.8	83.1	38.5	46.8	9.3	84.3	76.1
+deductive	63.5	86.6	40.4	48.6	11.4	85.8	74.1
+quasar	66.1	88.0	40.2	49.3	12.6	86.0	73.6
+Boethius$_{Unified}$	64.7	86.6	42.8	50.6	16.0	85.2	72.4
+Boethius$_{Decomposed}$	72.1	95.8	48.4	56.0	21.8	90.2	64.9
→ *SFT* (deductive)	67.3	85.4	49.3	53.9	22.9	85.0	63.9
→ *SFT* (Boethius)	76.1	97.5	54.7	62.3	28.7	96.0	58.3
→ *GRPO* (Boethius)	74.7	95.4	54.1	62.0	29.5	94.5	58.2
→ *SFT* + *GRPO* (Boethius)	76.8	97.5	56.2	65.4	34.8	96.2	54.5
Llama3-8B	66.1	90.2	42.0	51.0	11.5	90.6	73.2
+deductive	68.6	91.2	46.0	55.4	19.1	91.6	67.5
+quasar	69.6	92.0	47.2	57.2	22.0	92.4	65.4
+Boethius$_{Unified}$	70.6	94.4	46.8	60.2	28.0	92.4	63.6
+Boethius$_{Decomposed}$	80.2	96.2	63.0	66.4	35.6	93.7	50.7
→ *SFT* (deductive)	72.3	92.0	52.7	63.9	35.2	92.7	56.0
→ *SFT* (Boethius)	81.0	97.8	64.0	67.3	37.2	97.5	49.4
→ *GRPO* (Boethius)	79.2	96.1	62.3	65.5	35.1	96.0	50.4
→ *SFT* + *GRPO* (Boethius)	77.6	96.2	58.3	71.3	44.2	98.5	46.8
Llama3-70B	71.8	92.5	51.2	57.9	21.8	93.9	63.5
+deductive	73.0	93.7	52.2	60.6	25.4	95.8	61.2
+quasar	75.4	95.0	55.8	60.3	27.2	95.3	58.5
+Boethius$_{Unified}$	81.1	96.9	65.3	65.5	34.1	96.8	50.3
+Boethius$_{Decomposed}$	82.1	98.2	66.0	66.8	35.4	96.8	49.3
Gemma2-9B	67.2	84.1	50.4	57.9	24.1	91.6	62.8
+deductive	69.8	88.1	51.6	60.3	26.4	94.1	61.0
+quasar	70.5	89.7	52.0	60.8	27.0	94.6	60.5
+Boethius$_{Unified}$	69.8	89.5	46.0	60.2	28.5	92.0	62.8
+Boethius$_{Decomposed}$	72.6	90.4	54.8	62.8	31.8	94.7	56.7
→ *SFT* (deductive)	72.5	92.9	52.2	62.2	34.5	90.0	56.6
→ *SFT* (Boethius)	75.8	93.9	57.8	66.6	37.5	95.0	52.4
→ *GRPO* (Boethius)	74.2	91.4	57.0	66.9	36.5	95.4	53.2
→ *SFT* + *GRPO* (Boethius)	75.2	93.3	57.0	67.6	38.7	96.5	52.1
Qwen-3-1B	58.2	78.9	37.5	46.1	8.9	83.1	76.8
+deductive	61.9	84.3	39.5	48.2	11.0	85.4	74.8
+quasar	61.2	82.0	40.5	51.9	18.0	85.8	70.8
+Boethius$_{Unified}$	64.3	87.2	41.4	58.5	26.2	90.8	66.2
+Boethius$_{Decomposed}$	71.3	96.0	46.6	60.6	28.2	93.0	62.6
→ *SFT* (deductive)	71.0	89.0	52.6	57.2	22.0	92.5	62.7
→ *SFT* (Boethius)	76.6	97.9	55.4	62.2	30.3	96.2	57.1
→ *GRPO* (Boethius)	76.1	97.7	54.5	61.9	30.1	95.8	57.7
→ *SFT* + *GRPO* (Boethius)	78.7	98.7	58.8	65.2	34.2	96.2	52.5
Qwen3-8B	68.4	86.6	50.2	57.0	23.3	90.8	63.2
+deductive	66.2	88.5	44.2	59.3	24.6	94.0	65.6
+quasar	71.0	88.9	50.0	60.6	26.8	94.5	61.6
+Boethius$_{Unified}$	69.1	89.4	48.8	60.8	26.0	95.6	62.6
+Boethius$_{Decomposed}$	71.6	90.2	53.1	62.2	28.9	95.6	59.0
→ *SFT* (deductive)	69.5	90.4	48.0	63.1	32.0	94.3	57.6
→ *SFT* (Boethius)	74.3	92.2	57.4	66.2	36.6	95.8	53.0
→ *GRPO* (Boethius)	73.5	90.0	57.0	65.3	35.8	94.7	53.6
→ *SFT* + *GRPO* (Boethius)	74.8	92.7	57.0	66.5	35.2	96.2	50.1
Qwen3-32B	71.2	91.4	51.0	57.3	21.4	93.3	63.8
+deductive	72.8	93.7	51.8	60.3	25.2	95.4	61.5
+quasar	73.2	94.0	52.3	61.1	26.9	95.8	60.4
+Boethius$_{Unified}$	80.4	96.6	64.1	63.4	30.2	96.6	52.9
+Boethius$_{Decomposed}$	82.0	98.6	65.7	66.2	35.5	96.9	49.4
GPT-4o	73.4	93.5	53.3	58.9	23.1	94.7	61.8
+deductive	74.0	94.1	53.9	61.7	27.7	95.8	59.2
+quasar	76.0	95.3	55.0	62.3	28.3	96.3	58.4
+Boethius$_{Unified}$	80.7	96.9	64.9	63.4	28.9	97.9	53.1
+Boethius$_{Decomposed}$	81.9	97.7	66.2	65.7	33.4	98.0	50.2

Table 10: Results. Boethius$_{Unified}$ is schema-identification and schema-instantiation in a single prompt, Boethius$_{Decomposed}$ is schema-identification and schema-instantiation in 2 prompts under the same context.

G GRPO Reward Design

We use GRPO, which is guided by a composite reward function defined over model-generated reasoning traces. Each reward component is deterministic and verifiable, enabling stable optimisation without relying on learned reward models.
Let x denote the input and y a model-generated completion. The overall reward is defined as:

$$R(y) = r_{\text{s}}(y) + r_{\text{f}}(y) + r_{\text{SI}}(y) + r_{\text{abs}}(y). \quad (1)$$

1. Strict Answer Reward (r_{s}). This reward assesses syllogistic correctness by verifying whether the validity label extracted from the `<answer>` block matches the gold label $\hat{y}$.

$$r_{\text{s}}(y) = \begin{cases} 2 & \text{if } \texttt{extract\_label}(y) = \hat{y} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The extraction procedure is rule-based and fails if the answer block is missing/malformed.

2. Format Reward (r_{f}). To ensure adherence to the reasoning pipeline, we enforce a rigid, tagged output format. Each response is required to contain: `<schema>`, `<instantiation>`, and `<answer>`. It is verified using regular-expression matching and tag consistency checks. The reward is assigned as:

$$r_{\text{f}}(y) = \begin{cases} 0.5 & \text{if } \texttt{format\_valid}(y) = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

3. Structural Integrity Reward (r_{SI}). Beyond coarse-format validity, we encourage concise, well-formed reasoning traces by assigning incremental rewards to key structural markers and penalising excessive or irrelevant content.

$$r_{\text{SI}}(y) = \sum_{i=1}^{4} w_i \cdot \mathbb{1}(s_i \in y) - \lambda \cdot \textit{extra\_content}(y), \quad (4)$$

where s_i correspond to required structural elements within the reasoning trace (schema identifier, mapping, instantiated premises, decision), $w_i = 0.125$ are uniform weights, and $\lambda = 0.001$ penalises tokens generated outside the blocks.

4. Abstraction Reward (r_{abs}). To discourage surface-level reasoning and promote abstract manipulation of syllogistic schemas, we introduce an 'abstraction' reward that penalises lexical overlap between the input premises and the schema-level reasoning.
Let $V(x)$ denote the set of content-bearing tokens (e.g. nouns and adjectives) extracted from the input premises after lemmatisation and stopword removal, and let $T_{\text{schema}}(y)$ denote the tokens appearing within the `<schema>` block of the generated output. We define the normalised overlap as:

$$\text{overlap}(x, y) = \frac{|V(x) \cap T_{\text{schema}}(y)|}{|V(x)|}. \quad (5)$$

The abstraction reward is then computed as:

$$r_{\text{abs}}(y) = -\alpha \cdot \text{overlap}(x, y), \quad (6)$$

with $\alpha > 0$ a fixed scaling coefficient. Lower overlap corresponds to higher abstraction and is thus favoured during optimisation.
Abstraction Example. The abstraction reward is designed to prevent the model from encoding surface lexical content within the schema-level representation. Consider the syllogism: *All siameses are cats. All cats are felines.* A non-abstract schema-level response may restate the lexical content of the premises, e.g., "`Since all cats are felines and all siameses are cats, the conclusion follows.`" Such a response exhibits maximal lexical overlap with the input and is therefore penalised.
By contrast, an abstract schema-level response expresses the inference purely in symbolic terms, e.g., "`This instance instantiates the AA1 schema: All A are B; All B are C; therefore All A are C.`" In this case, no content-bearing tokens from the input premises appear in the schema block, resulting in zero lexical overlap and no abstraction penalty. This mechanism encourages the model to separate structural reasoning from lexical instantiation, thereby mitigating content effects and promoting symbol-like inference.

H Prompting Strategies

Strategy	Prompt Example
Deductive	Determine whether the following syllogistic argument is logically valid: *"All cats are mammals. Some cats are not pets. Therefore, some mammals are not pets."* Follow the instructions: 1. **Deductive Analysis:** Provide a step-by-step explanation assessing whether the conclusion follows logically from the premises. 2. **Final Label:** Output `VALID` or `INVALID`.
Schema Identification	Analyse the following syllogistic argument by abstracting away from content-specific meaning and considering only its formal structure: *"All cats are mammals. Some cats are not pets. Therefore, some mammals are not pets."* Follow the instructions: 1. **Identify Quantifier Pattern:** Describe the quantifiers and polarity of the premises and conclusion. 2. **Identify Figure:** Determine the structural arrangement of the terms (subject, middle, predicate). 3. **Output Schema:** Return the corresponding syllogistic schema label (as defined in Table 12).
Schema Instantiation	Given a syllogistic schema s identified in the previous step, evaluate whether the following argument correctly instantiates it: *"All cats are mammals. Some cats are not pets. Therefore, some mammals are not pets."* Follow the instructions: 1. **Instantiate Schema:** Map the argument's terms to the schema variables (A, B, C) and verify that the premises conform to the schema structure. 2. **Check Conclusion:** Determine whether the conclusion is licensed by the instantiated schema. 3. **Final Label:** Output `VALID` or `INVALID`.
Quasar	`Role` You are an experienced expert skilled in answering complex problems through logical reasoning and structured analysis. `Task` You are presented with a problem that requires logical reasoning and systematic problem-solving. Please answer the question following these steps rigorously: 1. Please consider the following question and exemplify the relevant predicates, variables, and constants. Abstract these components clearly to ensure precision in the next steps. Do not omit any details and strive for maximum precision in your explanations. Refer to this step as Abstraction (s1) 2. For each predicate, variable and constant defined in s1, translate the question into a formal symbolic representation. Ensure that the formalisation captures the question's logical structure and constraints. Provide the exact formalisation of each component exemplified in s1, referencing their corresponding definitions. Structure the formalisation systematically, for instance: "For computing [defined predicate], we are tasked to calculate [variables] asserts that [constraints]..". Refer to this step as Formalisation (s2) 3. Consider the formalisation in s2 in detail, ensure this is correct and solve the question by breaking down the steps operating a symbolic representation. Combine variables, constants, and logical rules systematically at each step to find the solution. Provide clear reasoning for each step. Structure the explanation systematically, for instance: "Step 1: Calculate... Step 2:....". Refer to this step as Explanation (s3) 4. In conclusion, behind explaining the steps supporting the final answer to facilitate the final evaluation, extract the answer in a short format by marking it as "The answer is ". At this stage, be strict and concise and refer to this step as Answering (s4). `Question` Determine whether the following syllogistic argument is logically valid, providing the label `Valid` or `Invalid`: *"All cats are mammals. Some cats are not pets. Therefore, some mammals are not pets."*

Table 11: Prompting Approaches.

I Syllogistic Schemas

Schema	Formulation		
	P_1	P_2	C
AA1	$\forall x(A(x) \to B(x))$	$\forall x(B(x) \to C(x))$	$\forall x(A(x) \to C(x))$
AA2	$\forall x(B(x) \to A(x))$	$\forall x(C(x) \to B(x))$	$\forall x(C(x) \to A(x))$
AA4	$\forall x(B(x) \to A(x))$	$\forall x(B(x) \to C(x))$	$\exists x(A(x) \wedge C(x))$
AI2	$\forall x(B(x) \to A(x))$	$\exists x(C(x) \wedge B(x))$	$\exists x(A(x) \wedge C(x))$
AI4	$\forall x(B(x) \to A(x))$	$\exists x(B(x) \wedge C(x))$	$\exists x(A(x) \wedge C(x))$
AO3	$\forall x(A(x) \to B(x))$	$\exists x(C(x) \wedge \neg B(x))$	$\exists x(C(x) \wedge \neg A(x))$
AO4	$\forall x(B(x) \to A(x))$	$\exists x(B(x) \wedge \neg C(x))$	$\exists x(A(x) \wedge \neg C(x))$
AE1	$\forall x(A(x) \to B(x))$	$\forall x(\neg C(x) \to \neg B(x))$	$\forall x(A(x) \to \neg C(x))$
AE2	$\forall x(B(x) \to A(x))$	$\forall x(C(x) \to \neg B(x))$	$\exists x(C(x) \wedge \neg A(x))$
AE3	$\forall x(A(x) \to B(x))$	$\forall x(C(x) \to \neg B(x))$	$\forall x(A(x) \to \neg C(x))$
AE4	$\forall x(B(x) \to A(x))$	$\forall x(B(x) \to \neg C(x))$	$\exists x(A(x) \wedge \neg C(x))$
IA1	$\exists x(A(x) \wedge B(x))$	$\forall x(B(x) \to C(x))$	$\exists x(A(x) \wedge C(x))$
IA4	$\exists x(B(x) \wedge A(x))$	$\forall x(B(x) \to C(x))$	$\exists x(A(x) \wedge C(x))$
IE1	$\exists x(A(x) \wedge B(x))$	$\forall x(\neg C(x) \to \neg B(x))$	$\exists x(A(x) \wedge \neg C(x))$
IE2	$\exists x(B(x) \wedge A(x))$	$\forall x(\neg C(x) \to \neg B(x))$	$\exists x(A(x) \wedge \neg C(x))$
IE3	$\exists x(A(x) \wedge B(x))$	$\forall x(\neg C(x) \to \neg B(x))$	$\exists x(A(x) \wedge \neg C(x))$
IE4	$\exists x(B(x) \wedge A(x))$	$\forall x(\neg C(x) \to \neg B(x))$	$\exists x(A(x) \wedge \neg C(x))$
OA3	$\exists x(A(x) \wedge \neg B(x))$	$\forall x(C(x) \to B(x))$	$\exists x(A(x) \wedge \neg C(x))$
OA4	$\exists x(B(x) \wedge \neg A(x))$	$\forall x(B(x) \to C(x))$	$\exists x(C(x) \wedge \neg A(x))$
EA1	$\forall x(A(x) \to \neg B(x))$	$\forall x(B(x) \to C(x))$	$\exists x(C(x) \wedge \neg A(x))$
EA2	$\forall x(B(x) \to \neg A(x))$	$\forall x(C(x) \to B(x))$	$\exists x(C(x) \wedge \neg A(x))$
EA3	$\forall x(A(x) \to \neg B(x))$	$\forall x(C(x) \to B(x))$	$\exists x(C(x) \wedge \neg A(x))$
EA4	$\forall x(B(x) \to \neg A(x))$	$\forall x(B(x) \to C(x))$	$\exists x(C(x) \wedge \neg A(x))$
EI1	$\forall x(A(x) \to \neg B(x))$	$\exists x(B(x) \wedge C(x))$	$\exists x(C(x) \wedge \neg A(x))$
EI2	$\forall x(\neg B(x) \to \neg A(x))$	$\exists x(C(x) \wedge B(x))$	$\exists x(C(x) \wedge \neg A(x))$
EI3	$\forall x(A(x) \to \neg B(x))$	$\exists x(C(x) \wedge B(x))$	$\exists x(C(x) \wedge \neg A(x))$
EI4	$\forall x(\neg B(x) \to \neg A(x))$	$\exists x(B(x) \wedge C(x))$	$\exists x(C(x) \wedge \neg A(x))$

Table 12: Complete list of the 27 classically valid categorical syllogistic schemas, expressed in predicate-logic form. The remaining 37 mood–figure combinations are classified as invalid and admit no valid conclusion (*nothing follows*), following the taxonomy by Bertolazzi et al. (2024).

J Example for an Unbelievable–Valid Syllogism

Task. *Determine whether the following syllogistic argument is logically valid.*

Argument.
Some siameses are cats.
No cats are felines.
Therefore, some siameses are not felines.

Baseline (Incorrect)

Judgement. The conclusion is rejected due to a conflict with common-sense world knowledge.

Output: INVALID
Failure mode: Immediate plausibility-based rejection.

Deductive Reasoning (Incorrect)

(Reasoning) Cats are commonly known to be felines, and siameses are a kind of cat. Since the second premise contradicts real-world knowledge, the argument is implausible. As a result, the conclusion does not meaningfully follow from the premises.

Output: INVALID
Failure mode: Content Effect (semantic plausibility overrides formal validity).

Boethius Unified — Partial Success (Incorrect)

Schema Identification. The argument instantiates the pattern: *Some A are B; No B are C; therefore Some A are not C*, corresponding to schema **IE1**.

Schema Instantiation (faulty). Because "cats" and "felines" are co-extensive in world knowledge, the negated relation in the second premise creates a contradiction. The instantiation is therefore rejected.

Output: INVALID
Failure mode: Correct schema identification, but instantiation contaminated by semantic knowledge.

Boethius Decomposed — Correct

Stage 1: Schema Identification. Quantifier structure: P1 = SOME(A, B); P2 = NO(B, C); C = SOME$(A, \neg C)$. This configuration matches the valid syllogistic schema **IE1**.

Stage 2: Schema Instantiation. Treating predicates as abstract placeholders: $A =$ *siameses*, $B =$ *cats*, $C =$ *felines*. From "Some A are B" and "No B are C", it follows that some A are not C. The conclusion is licensed by the schema independently of plausibility.

Output: VALID

Assessment. The baseline fails to discern implausibility from invalidity. The single-stage Boethius variant correctly recovers the schema but allows semantic leakage at the instantiation step. Only the dual-stage Boethius pipeline enforces a strict separation between form and content, yielding the correct validity judgement.

#Final Answer: VALID

Table 13: Comparison of deductive, Boethius single-stage reasoning, and Boethius dual-stage reasoning on an IE1 Unbelievable–Valid syllogism from GPT-4o generations.

K Example for an Unbelievable–Valid Syllogism (Second)

Task. *Determine whether the following syllogistic argument is logically valid.*

Argument.
Some sedans are cars.
No cars are vehicles.
Therefore, some sedans are not vehicles.

Baseline (Incorrect)

Judgement. The conclusion is rejected because the premises conflict with common-sense world knowledge.

Output: INVALID
Failure mode: Immediate plausibility-based rejection.

Deductive Reasoning (Incorrect)

Reasoning. Sedans are a type of car, and cars are universally understood to be vehicles. Since the second premise contradicts basic world knowledge, the argument is incoherent. Consequently, the conclusion cannot be logically supported.

Output: INVALID
Failure mode: Content Effect driven by strong taxonomic plausibility.

Boethius Unified — Partial Success (Incorrect)

Schema Identification. The argument follows the form: *Some A are B; No B are C; therefore Some A are not C*, which corresponds to the valid schema **IE1**.

Schema Instantiation (faulty). Because "cars" and "vehicles" are near-synonymous in common usage, the negation in the second premise invalidates the instantiation. The schema is therefore rejected as inconsistent with the argument content.

Output: INVALID
Failure mode: Correct schema recovery, but instantiation overridden by semantic equivalence.

Boethius Decomposed — Correct

Stage 1: Schema Identification. Quantifier Configuration: P1 = SOME(A, B); P2 = NO(B, C); C = SOME$(A, \neg C)$. This configuration uniquely matches the valid syllogistic schema **IE1**.

Stage 2: Schema Instantiation. Treating all predicates as abstract symbols: $A =$ *sedans*, $B =$ *cars*, $C =$ *vehicles*. Given "Some A are B" and "No B are C", it follows that some A are not C. The inference is licensed by the schema regardless of semantic plausibility.

Output: VALID

Assessment. The baseline exhibits immediate rejection driven solely by plausibility. Deductive reasoning amplifies this effect through explicit justification in terms of world knowledge. The single-stage Boethius variant correctly identifies the schema but fails to enforce abstraction during instantiation. Only the dual-stage Boethius pipeline consistently isolates formal structure from content, yielding the correct judgement.

#Final Answer: VALID

Table 14: Example from Qwen3-32B on an IE1 Unbelievable–Valid syllogism .