arXiv:2509.10222v1 [cs.AI] 12 Sep 2025

Compartmentalised Agentic Reasoning for Clinical NLI

Maël Jullien[1,3], Lei Xu [1,3], Marco Valentino[4], André Freitas[1,2,3]
[1]Department of Computer Science, University of Manchester, UK
[2] National Biomarker Centre, CRUK-MI, University of Manchester, UK
[3]Idiap Research Institute, Switzerland
[4] School of Computer Science, University of Sheffield, UK
[3]`{firstname.surname}@idiap.ch`

Abstract

A common assumption holds that scaling data and parameters yields increasingly structured, generalisable internal representations. We interrogate this assumption in clinical natural language inference (NLI) by adopting a benchmark decomposed into four reasoning families, *Causal Attribution*, *Compositional Grounding*, *Epistemic Verification*, and *Risk State Abstraction*, and introducing *CARENLI*, a Compartmentalised Agentic Reasoning for Clinical NLI that separates knowledge access from principled inference. CARENLI routes each premise–statement pair to a family-specific solver and enforces auditable procedures via a planner, verifier, and refiner.

Across four LLMs, CARENLI improves fidelity by up to 42 points, reaching 98.0% in *Causal Attribution* and 81.2% in *Risk State Abstraction*. Verifiers flag violations with near-ceiling reliability, while refiners correct a substantial share of epistemic errors. Remaining failures cluster in routing, identifying family classification as the main bottleneck. These results show that LLMs often retain relevant facts but default to heuristics when inference is underspecified, a dissociation CARENLI makes explicit while offering a framework for safer, auditable reasoning.

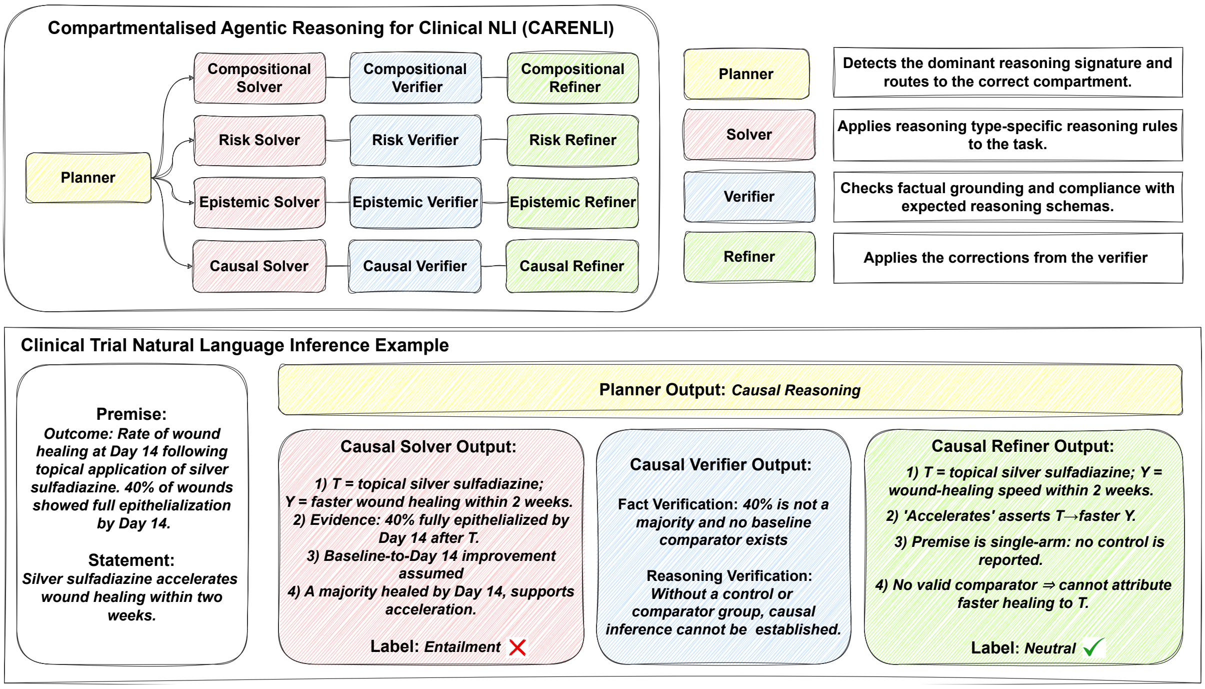
1 Introduction

Large language models (LLMs) achieve strong results across natural language processing benchmarks, often surpassing human baselines on standard natural language inference (NLI) datasets. Yet these successes mask systematic reasoning limitations, especially when tasks demand structured, domain-specific inference rather than surface-level pattern matching. In safety-critical contexts such as clinical research and practice, such failures are consequential: models may possess relevant knowledge but fail to apply it according to principled inferential rules, leading to unsafe or erroneous conclusions.

Clinical NLI exemplifies this challenge. Unlike general-domain NLI, which can often be solved by lexical alignment or frequency heuristics, clinical inference requires reasoning over causal attributions, multi-factor interactions, evidential conflicts, and latent risk states. Errors here have direct implications: conflating correlation with causation, overlooking toxic cross-factor regimens, deferring to unsupported authority, or ignoring catastrophic risks when severity outweighs frequency. Recent work therefore calls for systematic analysis of reasoning in clinical NLI, emphasising that success depends not on knowledge retrieval alone but on alignment with formalised reasoning flows.

We adopt and extend the Clinical Trial NLI (CTNLI) benchmark, which decomposes inference into four reasoning families: CAUSAL ATTRIBUTION, COMPOSITIONAL GROUNDING, EPISTEMIC VERIFICATION, and RISK STATE ABSTRACTION. Each family is associated with a formally specified schema, reflecting insights from cognitive psychology and formal semantics: counterfactual comparison, multi-factor constraint checking, evidence hierarchies, and severity–likelihood integration. Prior evaluations reveal a striking knowledge–reasoning dissociation: LLMs reliably encode clinical facts yet collapse into generic heuristics when reasoning, producing systematic rather than random errors.

To address this gap, we propose a compartmentalised agentic framework for clinical NLI, *Compartmentalised Agentic Reasoning for Clinical NLI (CARENLI)* (Figure 1), that routes each problem to its reasoning family. A planner identifies the family, a solver executes the inferential procedure, a verifier audits factual and structural validity, and a refiner enforces minimal corrections. By aligning each stage with the CTNLI schemas, the framework prevents heuristic shortcuts and enforces principled decision rules. We evaluate this approach across four contemporary LLMs—GPT-4o, GPT-



Planner	Detects the dominant reasoning signature and routes to the correct compartment.
Solver	Applies reasoning type-specific reasoning rules to the task.
Verifier	Checks factual grounding and compliance with expected reasoning schemas.
Refiner	Applies the corrections from the verifier

Premise: *Outcome: Rate of wound healing at Day 14 following topical application of silver sulfadiazine. 40% of wounds showed full epithelialization by Day 14.*

Statement: *Silver sulfadiazine accelerates wound healing within two weeks.*

Planner Output: *Causal Reasoning*

Causal Solver Output:

1) T = topical silver sulfadiazine; Y = faster wound healing within 2 weeks.
2) Evidence: 40% fully epithelialized by Day 14 after T.
3) Baseline-to-Day 14 improvement assumed
4) A majority healed by Day 14, supports acceleration.

Label: *Entailment* ✗

Causal Verifier Output:

Fact Verification: *40% is not a majority and no baseline comparator exists*

Reasoning Verification: *Without a control or comparator group, causal inference cannot be established.*

Causal Refiner Output:

1) T = topical silver sulfadiazine; Y = wound-healing speed within 2 weeks.
2) 'Accelerates' asserts T→faster Y.
3) Premise is single-arm: no control is reported.
4) No valid comparator ⇒ cannot attribute faster healing to T.

Label: *Neutral* ✓

Figure 1: CARENLI: Compartmentalised agentic reasoning framework for Clinical Trial NLI.

4o-mini, Gemini 2.5 Pro, and DeepSeek R1—under both reasoning-agnostic baselines and agentic conditions, including an oracle-planner ablation.

Our contributions are as follows: (i) CARENLI: a compartmentalised agentic reasoning framework for clinical NLI; (ii) substantial fidelity gains, with improvements up to +42 points over agnostic prompting and accuracies of 98.0% in causal attribution and 90.0% in compositional reasoning; (iii) enhanced reliability from verifier and refinement stages, with refinement correcting up to 39% of epistemic errors.

Together, these results demonstrate that structured, agentic prompting provides a principled pathway for aligning LLMs with domain-grounded inferential schemas, and more broadly highlight the importance of modular reasoning design for safety-critical applications.

2 Methodology

2.1 Tasks and Dataset

This evaluation is grounded in the domain of *clinical natural language inference* (CTNLI), where the objective is to determine whether a candidate *statement* is entailed, contradicted, or left undetermined by a given *premise*. A controlled benchmark introduced by Jullien et al. (2025) is adopted, organised into four reasoning families, each defined by explicit inferential criteria:

i. **Causal Attribution:** distinguishes between observational associations and true causal claims, requiring assessment of temporality, control conditions, and confounding.

ii. **Compositional Grounding:** evaluates whether the clinical validity of a statement can be inferred from structured configurations involving multiple interacting variables (e.g., dose, comorbidity, trial part).

iii. **Epistemic Verification:** tests the ability to evaluate the truth of a claim based on evidential support rather than speaker authority or assertion.

iv. **Risk State Abstraction:** examines reasoning about latent clinical risks and probable outcomes, particularly when these are not explicitly stated in the text.

The benchmark comprises ten CTNLI pairs per reasoning family, with each family formally specified through typed templates that constrain the inferential operations required for a correct decision. For example, causal attribution items require explicit comparison of event temporality and the presence of control groups, while risk abstraction items require mapping between latent risk factors and downstream consequences. In this way, each instance is anchored in a principled decision procedure, enabling precise diagnosis of reasoning failures. All items are instantiated from controlled parametric templates, ensuring that variation is restricted to clinically meaningful factors (e.g., patient characteristics, treatment regimens), while the underlying reasoning structure remains invariant.

To enhance robustness while maintaining diagnostic fidelity, the benchmark was extended by generating an additional ten instances per reasoning family using the same templates, resulting in 20 examples per family and 80 items in total.

2.2 Compartmentalised Reasoning

A central finding of Jullien et al. (2025) is that large language models exhibit a systematic *knowledge–reasoning dissociation*: they reliably encode the relevant clinical facts, yet fail to apply them in a manner consistent with principled inference. Performance on ground-knowledge probes approached ceiling, while accuracy on reasoning tasks collapsed. The errors were not stochastic but highly consistent, reflecting the application of uniform heuristics across different reasoning settings. Causal claims were reduced to temporal proximity, compositional interactions to single-attribute matches, epistemic evaluation to deference to authority, and latent risks to frequency counts. This pattern indicates that LLMs are not deficient in knowledge, but in their ability to *deploy* knowledge within appropriate reasoning schemas, manifested as single undifferentiated mode of inference, irrespective of the structural demands of the task.

From a cognitive perspective, this tendency toward monolithic reasoning contrasts sharply with human inference. Human reasoning is well established is ungoverned by a uniform process, but instead flexibly recruits distinct inferential strategies, adapting the method of reasoning to the structure and demands of the problem. Counterfactual simulation underlies causal attribution, constraint satisfaction supports compositional reasoning, evidential coherence governs epistemic verification, and probabilistic projection guides risk estimation (Sloman and Sloman, 2009; Kahneman, 2011; Johnson-Laird, 1983; Tversky and Kahneman, 1974). Each strategy depends on specialised representational resources and inferential checks, and they are not interchangeable. LLMs fail precisely because they do not distinguish among these strategies, applying a single heuristic mode where human cognition would adaptively compartmentalise.

From a semantic perspective, the four reasoning families in CTNLI correspond to qualitatively distinct relational structures, each governed by its own logical or probabilistic rules. Causal attribution depends on interventionist counterfactuals (Pearl, 2009; Lewis, 1973), compositional grounding on multi-factor compatibility constraints, epistemic verification on evidential hierarchies, and risk abstraction on the integration of severity and probability (Van Fraassen, 1980). Collapsing these type-distinct relations into a single generic reasoning process erases the inferential distinctions that render them meaningful. The systematic heuristics observed in LLM outputs can thus be understood as the consequence of ignoring type distinctions that formal semantics treats as foundational.

Taken together, these findings motivate the need for compartmentalised reasoning in CTNLI. If failures arise because LLMs apply a uniform heuristic mode of inference to semantically heterogeneous tasks, then improving performance requires enforcing the separation of reasoning families and constraining inference to the principles specific to each. Compartmentalisation addresses precisely this gap: it reinstates the cognitive and semantic boundaries that distinguish causal, compositional, epistemic, and risk-based inference, preventing models from collapsing them into a single generic process.

Reasoning-agnostic prompting, whether through direct answering or free-form CoT, provides no such distinction and therefore encourages heuristic shortcuts that ignore the formal criteria of each reasoning type. By contrast, a compartmentalised design ensures that tasks are first recognised as belonging to a specific reasoning family, and are then solved by following the schema formally associated with that family. This approach not only aligns with evidence from cognitive psychology and formal semantics but also creates a practical mechanism for improving reasoning fidelity: once compartmentalised, errors can be attributed to either factual inaccuracy or violation of the family-specific schema, allowing for principled verification and correction.

In the next section, we instantiate this principle through *CARENLI* (Figure 1), in which inference is decomposed into specialised roles: a planner that identifies the reasoning family, a solver that applies the appropriate decision procedure, a verifier that audits factual and structural validity, and a refiner that enforces minimal corrections.

2.3 Compartmentalised Agentic Reasoning for Clinical NLI (CARENLI)

An *agentic* framework provides a natural instantiation of compartmentalised reasoning. By decomposing inference into interacting but specialised agents, the system enforces separation of functions while retaining the capacity for coordination. This design aligns directly with the requirements of clinical NLI, where reasoning must follow family-specific inferential principles: distinct reasoning families can be assigned to specialised agents, whose outputs can then be subject to explicit verification and refinement. Moreover, the modularity of the approach permits straightforward extension to additional forms of formalised reasoning beyond those represented in this dataset, as new agents can be introduced to capture alternative inferential patterns while preserving the same overarching pipeline.

The resulting CARENLI framework consists of four roles: a **Planner** that identifies the relevant reasoning family, a **Solver** that executes the corresponding inference procedure, a **Verifier** that audits factual and structural validity, and a **Refiner** that applies minimal corrections, as shown in Figure 1. Each agent is constrained to a specific reasoning function, ensuring that the overall system embodies the principle of compartmentalisation while retaining the flexibility of multi-agent coordination.

Planner. The planner implements compartmentalisation at the level of *problem recognition*. For each premise–statement pair, it selects exactly one of the four reasoning families—CAUSAL ATTRIBUTION, COMPOSITIONAL GROUNDING, EPISTEMIC VERIFICATION, or RISK STATE ABSTRACTION—and outputs this as a routing signal. The planner does not infer truth values; its role is restricted to identifying the dominant inferential signature. The planner identifies the *root signature* and resolves ambiguity through a fixed hierarchy, ensuring that instances with multiple inferential elements are routed to the most structurally demanding family.

To achieve this, the planner is prompted with concise definitions of the four families, each specified by its inferential signature: causal attribution by intervention–outcome claims, compositional grounding by joint clinical constraints, epistemic verification by the adjudication of conflicting evidence, and risk abstraction by relative or latent risk comparisons.

2.4 Solver Specifications

Solver. The solver operationalises *family-specific inference* by combining the *core principles* of each reasoning type with its associated *decision procedure*. Core principles define the normative criteria of validity, such as the requirement for comparators in causal inference, compatibility constraints in compositional reasoning, evidential hierarchies in epistemic verification, or severity–likelihood integration in risk abstraction. The decision procedure specifies how these principles are applied step by step to premise–statement pairs, constraining the solver to produce outcomes consistent with the intended reasoning pathway. By enforcing this alignment, the solver prevents collapse into generic heuristics and ensures judgments reflect the formal inferential schemas.

Epistemic Verification Solver. The Epistemic Verification solver operationalises verification when premises contain multiple, potentially conflicting assertions. Following Jullien et al. (2025), premises are treated as conjunctions of defeasible epistemic commitments $K_a(\varphi_i)$, each attributed to a source agent a (Fagin et al., 2004; Van Benthem, 2011). Three principles govern inference. First, all assertions are treated as reported claims rather than ground truth, with admissibility evaluated against the clinical model $\mathcal{M}\text{CT} = \langle W\text{CT}, I_{\text{CT}} \rangle$. Second, inconsistencies are resolved using a plausibility function $\pi(\varphi_i)$ instantiated as an evidential hierarchy: objective measurements dominate diagnostic criteria, which dominate observations, interpretations, and self-report. Conflicts are resolved by discarding lower-ranked commitments, yielding a maximal consistent set $E^* \subseteq E$ such that $\forall w \in W_{\text{CT}}, E^*$ is jointly satisfiable under $\mathcal{M}_{\text{CT}}$. Third, coherence constraints exclude ontologically impossible or temporally incoherent assertions. A candidate statement s is entailed if $E^* \models s$, contradicted if $E^* \models \neg s$, and neutral if neither holds. Neutrality is therefore a principled outcome when conflicts remain unresolved or evidence underdetermined

Causal Attribution Solver. The Causal Attribution solver operationalises attribution when the hypothesis asserts a treatment → outcome relation. Under the interventionist framework (Pearl, 2009), causal effect is defined as

$$\text{CE}(T, Y) \triangleq \mathbb{E}[Y \mid \text{do}(T = 1)] - \mathbb{E}[Y \mid \text{do}(T = 0)],$$

where T denotes a treatment, Y an outcome, and $\mathrm{do}(T = t)$ an intervention severing incoming dependencies. Unlike observational associations $\mathbb{E}[Y \mid T = t]$, causal effect requires interventional contrast. The solver enforces this distinction by (i) parsing premises for outcome measurements or adverse events, treated as observational unless comparators or manipulations are explicit; (ii) verifying causal criteria of temporality, contrast, and confounding control; and (iii) evaluating the hypothesis s: entailment requires interventional evidence, contradiction follows from unsupported causal claims, and neutrality applies otherwise. This constrains models to uphold the semantics of interventionist causality and corrects the failure mode observed in Jullien et al. (2025), where models collapsed causal attribution into association, predicting entailment from surface correlations rather than from interventional criteria.

Compositional Grounding Solver. The Compositional Grounding operationalises inference when the truth of a statement depends on the joint configuration of multiple clinical factors rather than any single predicate. Formally, compositional grounding requires that a tuple

$$x := \langle d, z, dx, s \rangle \in \mathfrak{D}_{\mathrm{CT}}^4$$

of drug d, dose z, diagnosis dx, and schedule s must be admissible under the interpretation function I_{CT}. A statement ψ asserting clinical benefit is entailed only if $I_{\mathrm{CT}}(\mathrm{Benefit})(x) \to \mathit{True}$ in all admissible worlds $w \in W_{\mathrm{CT}}$; contradictions arise when x violates therapeutic, diagnostic, or scheduling constraints. First, atomic factors (drug, dose, schedule, patient details) are extracted. Second, the tuple x is assembled and tested for compatibility under I_{CT}, which encodes therapeutic ranges, scheduling rules, and indication–diagnosis mappings. Third, the hypothesis s is assessed: entailment requires that x be admissible and support the asserted outcome; contradiction arises when x violates constraints; neutrality is assigned when the configuration is underdetermined. This design intends to prevents the error mode identified in Jullien et al. (2025), where LLMs collapsed compositional interactions into isolated surface predicates, thereby overlooking protocol violations or emergent incompatibilities.

Risk Abstraction Solver. The Risk Abstraction solver operationalises inference when hypotheses involve explicit or latent risk. Risk is defined as an expectation over admissible worlds $w \in W_{\mathrm{CT}}$, combining probability and severity of adverse events (Hunink et al., 2014):

$$\mathbb{E}_{w \sim \Pr(w|\varphi)} \left[\sum_{e \in \mathcal{E}(w,\psi)} \Pr(e \mid w) \cdot \mathcal{A}(e, w) \right],$$

where $\mathcal{E}(w, \psi)$ is the set of clinically relevant events for ψ, $\Pr(e \mid w)$ their conditional probability, and $\mathcal{A}(e, w)$ an adverse outcome function quantifying harm. Entailment requires that the statement correctly reflect this expected risk profile, either by ranking events appropriately (when explicit frequencies are given) or by acknowledging unruled-out harms in latent risk settings.

The solver prompt translates this into a decision procedure with three steps. First, it identifies adverse events or conditions in the premise and classifies them according to severity and probability. Second, it integrates these two dimensions into an expected-harm ranking: severe but rare outcomes may outweigh frequent but benign ones, in line with clinical reasoning. Third, it evaluates the hypothesis s: entailment follows if s reflects the correct harm-weighted risk ordering, contradiction if it inverts or ignores it, and neutrality if the evidence base is insufficient.

By enforcing this expected-risk computation, the solver ensures that judgments reflect decision-theoretic reasoning rather than surface counts or lexical salience. This is intended to correct the failure mode highlighted in Jullien et al. (2025), where models frequently misclassified risk when severity and frequency were in tension, defaulting to naive frequency matching rather than principled risk abstraction (Hunink et al., 2014; Eisenhauer et al., 2009; CTC, 2017).

Verifier. The verifier operationalises *auditable reasoning* by checking solver outputs for factual fidelity and adherence to family-specific protocols. Two checks are applied. First, **fact verification** ensures all knowledge invoked is either premise-grounded or admissible as general clinical regularity, filtering out unsupported or fabricated content (Gravel et al., 2023; Aljamaan et al., 2024). Second, **pattern verification** enforces alignment with the expected reasoning patterns, flagging and minimally correcting deviations from family-specific inferential schemas. Prior work demonstrates that such refinement layers enhance reliability in LLM reasoning by correcting shallow or inconsistent outputs and anchoring them to explicit procedural rules (Webson and Pavlick, 2022; Quan et al., 2024). This design is supported by the experimental findings of Jullien et al. (2025), which demonstrated that LLMs generally encode the relevant clinical facts, and are able to recognise both expected, and degenerate reasoning, but systematically fail to apply this knowledge during inference.

Refiner. The refiner constitutes the final stage, implementing corrections prescribed by the verifier. It does not generate new inferences but minimally edits solver outputs to restore conformity. First, it integrates fact verification, removing or replacing unsupported claims with premise-grounded or admissible ones. Second, it applies pattern verification, restructuring reasoning to align with the family-specific decision schema—for instance, enforcing comparator requirements in causal inference (Pearl, 2009) or harm-weighted reasoning in risk assessment (Hunink et al., 2014). Finally, the refiner produces a corrected reasoning trace and final entailment judgment, aligned with the decision protocol and reasoning principles of its corresponding solver. In this way, the refiner ensures that outputs reflect both factual validity and principled reasoning, consistent with evidence that refinement stages improve LLM reliability by anchoring them to explicit rules (Webson and Pavlick, 2022; Quan et al., 2024).

3 Empirical Evaluation

We evaluate the *CARENLI* framework across four contemporary large language models: GPT-4o and GPT-4o-mini (Hurst et al., 2024), Gemini 2.5 Pro (Comanici et al., 2025), and DeepSeek R1 (Guo et al., 2025). These models were selected to represent diverse architectural families, training regimes, and performance tiers.

Three evaluation settings are considered. First, the *CARENLI* framework, in which all agents are active. Second, an *Oracle* CARENLI ablation, in which the planning stage is bypassed and the correct reasoning family is supplied directly to downstream agents. Third, baseline prompting conditions, in which models are tested under CoT prompting and direct answering (Agnostic CoT, and Agnostic Direct). this setting, prompts are reasoning-type agnostic and provide no guidance toward the family-specific inferential schemas.

Each model–strategy pair is evaluated over the full dataset of 80 examples (20 per reasoning family), with four independent runs per configuration.

4 Results

4.1 Overall Performance Patterns

***CARENLI* achieves near-ceiling causal reasoning and major gains in risk abstraction** The *CARENLI* framework yields substantial improvements in reasoning fidelity across all families (Figure 2, and Table 1). *CARENLI* accuracy consistently exceeds 70% in three of the four reasoning types, with peak performance reaching 98.0% in CAUSAL ATTRIBUTION and 81.2% in RISK STATE ABSTRACTION. Even the more challenging families, such as EPISTEMIC VERIFICATION, achieve accuracies around 76–80%, substantially higher than in prior baseline conditions. The only persistent weakness is COMPOSITIONAL GROUNDING, where accuracy remains below 50% under *CARENLI* routing.

Residual errors in compositional and risk reasoning stem from misclassification, not solver limits The Oracle Planner condition confirms that these residual errors are primarily attributable to misclassification rather than solver incapacity. When supplied with the correct reasoning family, accuracy in COMPOSITIONAL GROUNDING rises dramatically, from 18–48% under *CARENLI* routing to 90% in the best case. Similar gains are observed in RISK STATE ABSTRACTION, where performance reaches 100% under Oracle Planner evaluation. These results demonstrate that models are capable of following the expected inferential schemas once correctly routed, and that the principal bottleneck lies in classification accuracy rather than in the execution of reasoning procedures.

Taken together, these findings establish that large language models benefit systematically from compartmentalisation. Whereas uniform prompting encourages a single heuristic mode of inference, *CARENLI* enforces adherence to family-specific schemas, producing accuracy improvements of up to 40 percentage points and near-ceiling performance in causal reasoning. The central conclusion is that reasoning fidelity is constrained not by the expressive capacity of LLMs, but by whether inference is decomposed into type-specific compartments aligned with principled decision rules.

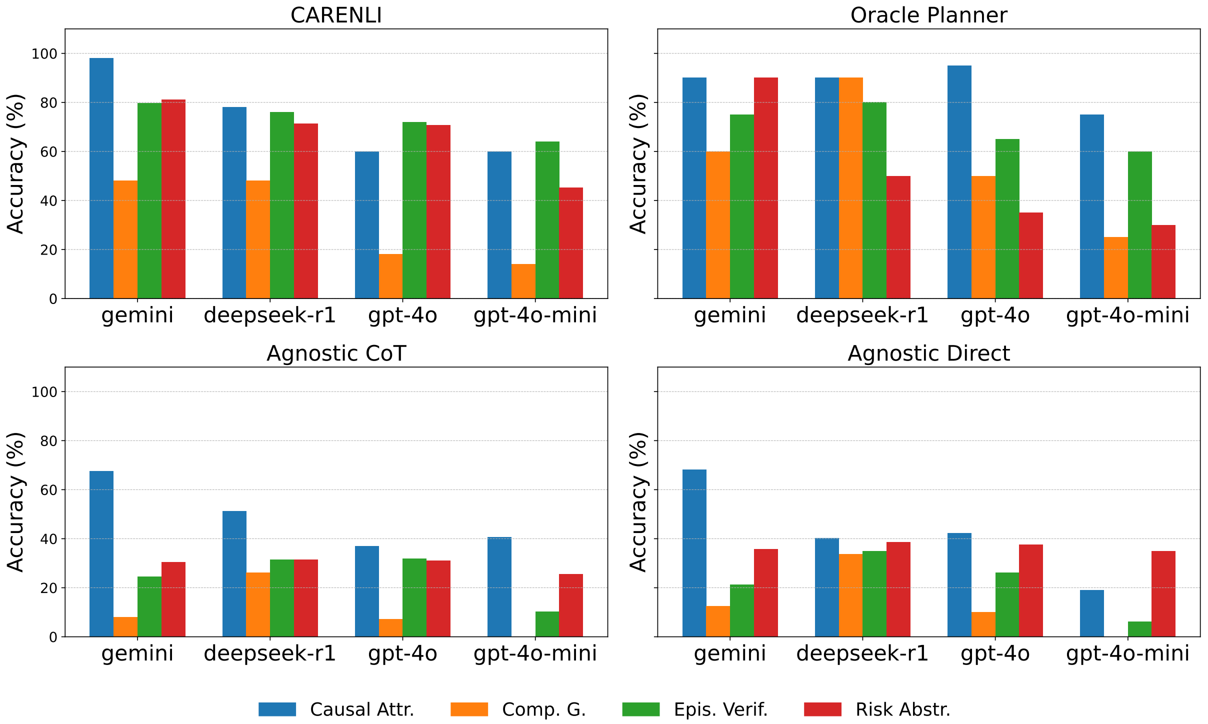
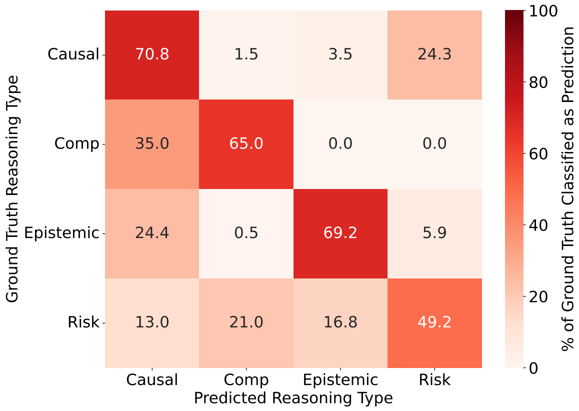


Figure 2: Overall accuracy on CTNLI tasks across all models and evaluation strategies (*CARENLI*, *Oracle Planner*, *CoT*, and *Direct*). Results are averaged over four runs per configuration.



	Causal	Comp	Epistemic	Risk
Causal	70.8	1.5	3.5	24.3
Comp	35.0	65.0	0.0	0.0
Epistemic	24.4	0.5	69.2	5.9
Risk	13.0	21.0	16.8	49.2

Figure 3: Confusion matrix of *CARENLI* reasoning family classification accuracy.

4.2 Reasoning Classification Accuracy

Classification accuracy is uneven across families, strongest in causal attribution Reasoning family classification is only moderately reliable (Figure 4, and Table 2), with accuracies ranging from 36.0% to 92.2% across families. CAUSAL ATTRIBUTION is most consistently identified, with performance above 54% across all systems, whereas COMPOSITIONAL GROUNDING remains persistently difficult, falling to 36.0% in some conditions. EPISTEMIC VERIFICATION is classified at moderate levels (67–72%), and RISK STATE ABSTRACTION is weakest overall, typically between 36.7% and 60.2%.

Models systematically misroute risk and epistemic tasks as causal The confusion matrix (Figure 3) shows that these errors are systematic rather than random. CAUSAL ATTRIBUTION is overpredicted, absorbing nearly one quarter of risk items and one fifth of epistemic items. This reflects a tendency of models to default to causal schemas whenever causal language is present. In contrast, RISK STATE ABSTRACTION is rarely overpredicted but frequently under-classified, often misrouted to causal or epistemic categories. These biases confirm that models lack robust reasoning-type discrimination, instead applying a monolithic causal heuristic across diverse problems.

4.3 Effect of Classification on Task Accuracy

Correct routing yields high solver accuracy, misclassification causes severe degradation Task success is tightly coupled to classification accuracy. When items are routed correctly, solvers achieve strong performance, frequently exceeding 70% and reaching near-ceiling in CAUSAL ATTRIBUTION (98.0%, Figure 5, and Table 1). Misclassification, however, leads to severe degradation: COMPOSI-

Model	Causal Attr.	Comp. G.	Epis. Verif.	Risk Abstr.
CARENLI				
gemini	**98.0**	**48.0**	**79.7**	**81.2**
deepseek-r1	78.0	**48.0**	76.0	71.3
gpt-4o	60.0	18.0	72.0	70.7
gpt-4o-mini	60.0	14.0	64.0	45.3
Oracle Planner				
gemini	90.0	60.0	75.0	90.0
deepseek-r1	90.0	**90.0**	80.0	**50.0**
gpt-4o	**95.0**	50.0	65.0	35.0
gpt-4o-mini	75.0	25.0	60.0	30.0
Agnostic CoT				
deepseek-r1	51.2	26.3	**31.5**	**31.5**
gemini	**67.7**	8.0	24.5	30.5
gpt-4o	37.0	7.2	**32.0**	31.2
gpt-4o-mini	40.7	0.0	10.3	25.5
Agnostic Direct				
deepseek-r1	40.3	33.8	35.0	38.7
gemini	**68.2**	12.5	21.3	35.7
gpt-4o	42.3	10.0	26.2	37.7
gpt-4o-mini	19.0	0.0	6.2	35.0

Table 1: Accuracy across reasoning tasks.

Model	Causal Attr.	Comp. G.	Epis. Verif.	Risk Abstr.
gemini	**84.7**	**92.2**	**72.3**	**60.2**
deepseek-r1	80.0	56.0	66.7	48.7
gpt-4o	64.0	74.0	67.3	52.0
gpt-4o-mini	54.0	36.0	70.7	36.7

Table 2: Reasoning Classification Accuracy across tasks

tional Grounding drops below 50% despite classification accuracies exceeding 90%, and Risk State Abstraction falls from 81.2% under correct routing to as low as 45.3% when misclassified.

Classification–accuracy correlation is strongest in causal and risk reasoning, weakest in epistemic This dependency is quantified by Spearman correlations between classification and task accuracy. Across all families, the correlation is moderate ($\rho = 0.38$), but strongly family-dependent: near-deterministic in Causal Attribution ($\rho = 0.95$) and Risk State Abstraction ($\rho = 0.80$), moderate in Compositional Grounding ($\rho = 0.63$), and minimal in Epistemic Verification ($\rho = 0.20$). These figures confirm that causal and risk reasoning require correct routing, while epistemic reasoning shows partial robustness because it can be approximated by surface heuristics.

Classification reliability is the central bottleneck of the framework Overall, classification emerges as the central bottleneck of the framework. Solvers can apply family-specific schemas when routed correctly, but performance collapses under misclassification, with compositional reasoning uniquely brittle and epistemic reasoning only partially resilient.

4.4 Framework versus CoT and Direct Prompting

Compartmentalisation improves accuracy by 30+ points over CoT and Direct prompting Compartmentalisation provides a systematic advantage over CoT and Direct prompting. Across reasoning families, the full framework exceeds baseline performance by margins of 30 points or more. For example, in Risk State Abstraction accuracy rises from 25–39% under CoT/Direct to 71–81% in the framework, and in Causal Attribution from 37–68% to 78–98%. These improvements hold across models, underscoring that the effect is not system-specific but reflects a general property of structured prompting.

Even misclassified solver prompts outperform CoT and Direct baselines Importantly, even when classification is incorrect, solvers guided by

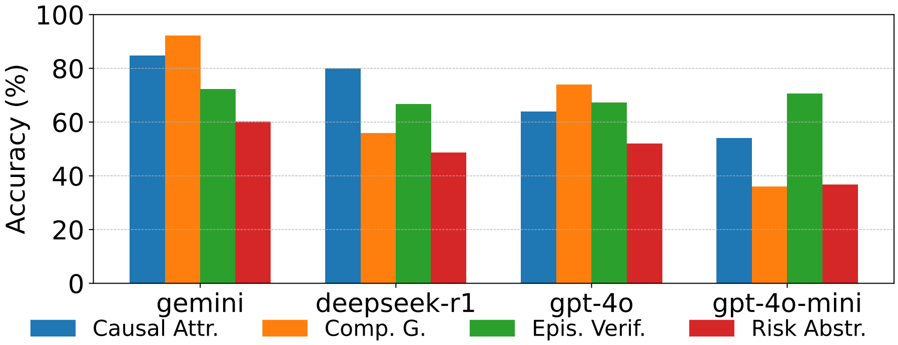


Figure 4: Overall accuracy of reasoning classification across models.

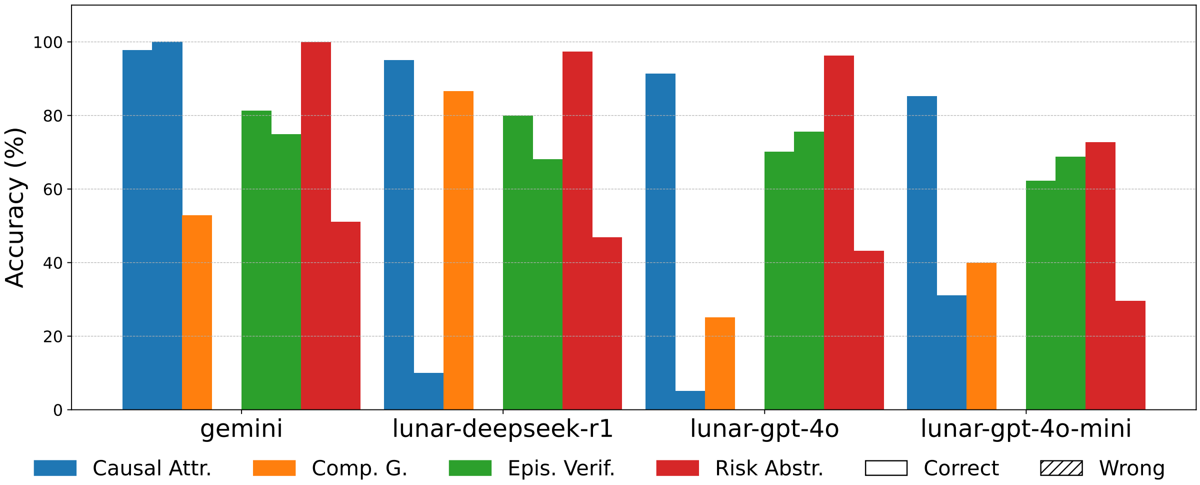


Figure 5: Accuracy on reasoning tasks when classification of the reasoning type was correct versus incorrect.Results are averaged per reasoning family across models.

family-specific instructions still outperform CoT and Direct (Figure 5). In EPISTEMIC VERIFICATION, misclassified solver prompts achieve 68–75% accuracy, compared to only 21–32% under baseline prompting. This resilience arises because solver prompts encode structural constraints—such as comparator requirements, evidential hierarchies, or severity–likelihood integration—that continue to shape reasoning even under incorrect routing.

Oracle Planner confirms solvers can follow schemas when routed correctly The Oracle Planner condition further confirms that residual errors stem primarily from classification. Supplying the correct reasoning family raises accuracy in COMPOSITIONAL GROUNDING from below 50% to 90% and in RISK STATE ABSTRACTION to 100%. These findings establish that solvers can implement the expected reasoning schemas, and that the absence of reliable routing is the key barrier to performance under the full framework.

4.5 Verifier Performance

Verifier judgments frequently exceed solver accuracy across families The verifier stage provides an independent audit of solver outputs, checking both factual grounding and adherence to reasoning schemas. Verifier accuracy is measured by whether the verifier correctly flags solver reasoning as invalid whenever the solver's initial label prediction is incorrect, and conversely affirms correctness when the solver's label is correct. Verifier accuracy frequently exceeds solver accuracy itself (Figure 6, Table 3). In CAUSAL ATTRIBUTION, verifier accuracy reaches 98–100%, while in RISK STATE ABSTRACTION it ranges from 71–100%, with perfect performance achieved in some Oracle Planner runs. By contrast, COMPOSITIONAL GROUNDING remains the most difficult to verify, with accuracy between 20–60%.

Verification is easier in causal and risk reasoning, harder in composition Two conclusions fol-

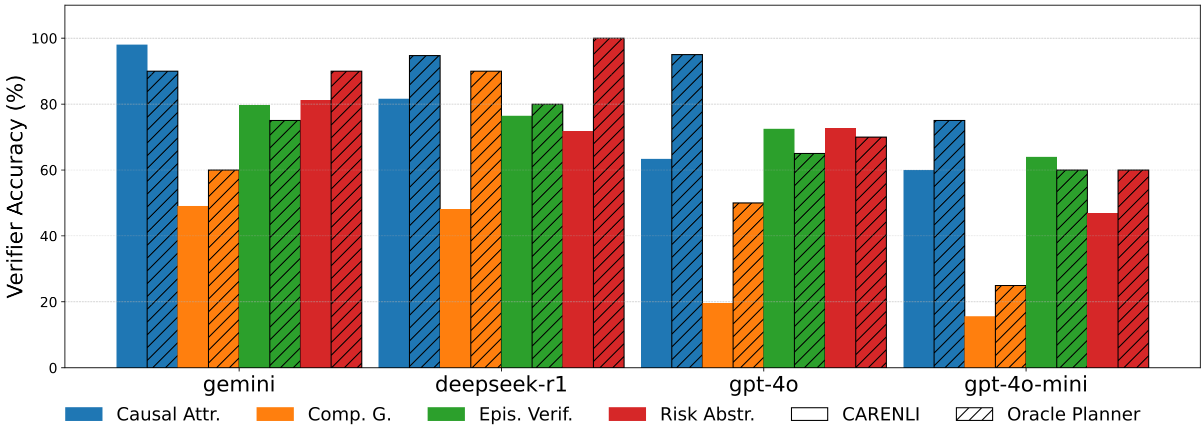


Figure 6: Verifier accuracy across reasoning families and models.

Model	Causal Attr.	Comp. G.	Epis. Verif.	Risk Abstr.
CARENLI				
gemini	**98.0**	**49.1**	**79.7**	**81.2**
deepseek-r1	81.6	48.0	76.5	71.8
gpt-4o	63.4	19.7	72.5	72.6
gpt-4o-mini	60.0	15.6	64.0	46.9
Oracle Planner				
gemini	90.0	60.0	75.0	90.0
deepseek-r1	94.7	**90.0**	**80.0**	**100.0**
gpt-4o	95.0	50.0	65.0	70.0
gpt-4o-mini	75.0	25.0	60.0	60.0

Table 3: Verifier accuracy (%) across reasoning tasks

low. First, LLMs are more reliable at recognising invalid reasoning than at producing valid reasoning directly: even when solver accuracy falls below 50%, verifier judgments often exceed 70%. Second, verification is easiest in families with strong structural signals, such as temporal precedence in causality or severity–likelihood trade-offs in risk, and most difficult in composition, where violations depend on multiple interacting factors. These results confirm that while models struggle to apply schemas, they can reliably detect their violation, making verification a robust complement to compartmentalisation.

4.6 Effects of Refinement

Refinement corrects up to 39% of epistemic errors through factual edits The refinement stage applies minimal edits to solver outputs in order to correct factual inaccuracies and restore alignment with family-specific reasoning schemas. Its impact is heterogeneous across reasoning families and models (Figure 7, and Table 4). In some settings, refinement yields clear improvements: in EPISTEMIC VERIFICATION, accuracy increases by up to 39 percentage points, with more than half of revisions correcting initial errors (Figure 7, and Table 5).

Refinement is rare but highly accurate in causal reasoning, common but ineffective in composition In other families, however, refinement delivers limited gains. In CAUSAL ATTRIBUTION, revisions are infrequent, typically affecting fewer than 10% of predictions (e.g., 6.0% for GPT-4o under *CARENLI*, 10.0% under Oracle Planner), but they are highly effective: nearly all changes corrected the initial label (100% gains). By contrast, in COMPOSITIONAL GROUNDING, revisions are far more common, with change rates of 22–35% depending on the model, yet accuracy rarely improves, with gains observed in only 16–29% of those cases. This pattern reflects the brittleness of compositional reasoning: once a multi-factor configuration is misconstructed, local corrections cannot repair the violation.

Fact errors are easier to fix than reasoning errors, limiting compositional gains A finer-grained error analysis clarifies these dynamics. In

Model	Causal Attr.	Comp. G.	Epis. Verif.	Risk Abstr.
CARENLI				
gemini (init.)	98.0	49.1	79.7	81.2
gemini (refined)	98.0	48.0	79.7	81.2
deepseek-r1 (init.)	81.6	48.0	76.5	71.8
deepseek-r1 (refined)	78.0	48.0	76.0	71.3
gpt-4o (init.)	63.4	19.7	72.5	72.6
gpt-4o (refined)	60.0	18.0	72.0	70.7
gpt-4o-mini (init.)	60.0	15.6	64.0	46.9
gpt-4o-mini (refined)	60.0	14.0	64.0	45.3
Oracle Planner				
gemini (init.)	90.0	60.0	75.0	90.0
gemini (refined)	90.0	60.0	75.0	90.0
deepseek-r1 (init.)	94.7	90.0	80.0	100.0
deepseek-r1 (refined)	90.0	90.0	80.0	50.0
gpt-4o (init.)	95.0	50.0	65.0	70.0
gpt-4o (refined)	95.0	50.0	65.0	35.0
gpt-4o-mini (init.)	75.0	25.0	60.0	60.0
gpt-4o-mini (refined)	75.0	25.0	60.0	30.0

Table 4: Accuracy of initial predictions vs refined predictions

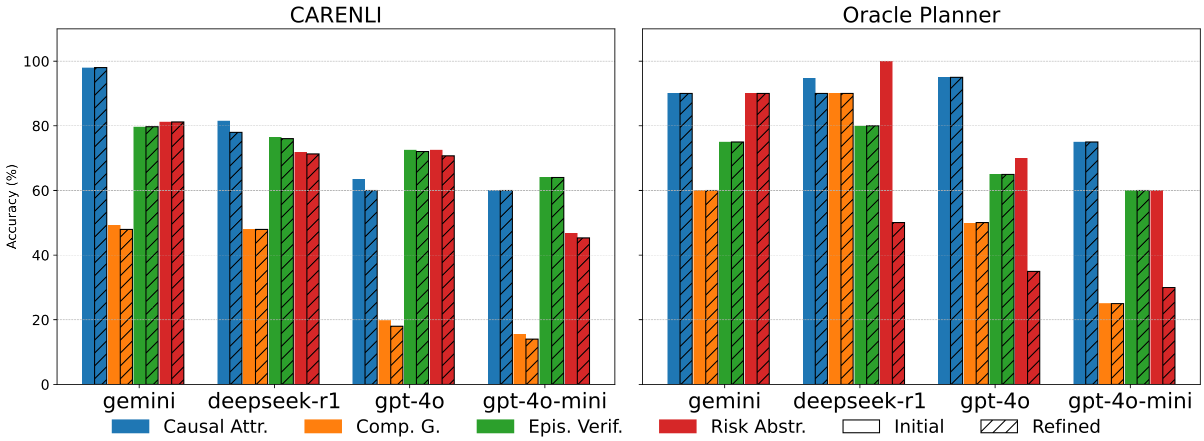


Figure 7: Comparison of solver outputs before and after refinement across reasoning families.

Epistemic Verification, revisions occur in up to 39.3% of cases (GPT-4o-mini), with over half of changes correcting the label (54–73% gains), as unsupported claims can be removed or replaced with premise-grounded evidence. In Risk State Abstraction, revisions are less frequent (5–16.6%) and only partially successful (0–60% gains), consistent with the difficulty of re-estimating severity–likelihood trade-offs post hoc. Overall, fact errors are easier to correct than reasoning errors, yet the latter predominate: reasoning violations accounted for 68.1% of cases, compared to 61.9% involving unsupported or fabricated facts. Because refinement operates through minimal local corrections, it can reliably address factual inaccuracies but rarely resolves structural reasoning failures. This asymmetry explains why refinement improves accuracy in epistemic and causal reasoning, but fails to salvage compositional tasks.

In sum, refinement is most effective for correcting local factual inaccuracies but limited against deeper reasoning errors. These results reinforce the knowledge–reasoning dissociation (Jullien et al., 2025): models typically possess relevant facts and can identify errors, yet fail to reliably apply domain knowledge within principled reasoning flows. Refinement partially mitigates this by enforcing factual and procedural alignment, but cannot resolve structural inferential deficiencies.

4.7 Qualitative Analysis of Representative Items

A qualitative analysis of four representative cases, one from each reasoning family, is presented using outputs from DeepSeek R1. For each case, the reasoning trace generated under the *oracle plan-*

ner condition (agentic framework, with the solver routed to the correct reasoning family) is contrasted with the baseline produced under *agnostic CoT* prompting. This comparison highlights how explicit compartmentalisation enables alignment with the expected reasoning patterns specified in the knowledge framework, while also revealing the characteristic deviations of free-form CoT.

Causal Attribution (Problem 6).

> **Premise:** *"Adverse Events Summary: Total: 4/12 (33.33%). Headache 2/12 (16.67%), Pruritus 1/12 (8.33%), Mild anemia 1/12 (8.33%). All events were Grade 1 or 2."*
> **Statement:** *"The drug was effective and well tolerated, with only mild side effects reported."*
>
> **Agentic (verbatim).** *''The statement makes two claims: (1) effectiveness ($T \rightarrow$ improved outcome) and (2) tolerability ($T \rightarrow$ only mild side effects). The premise only reports adverse event grades (all Grade 1/2) and rates (33.33% total), which directly confirms tolerability. However, the premise contains no efficacy data (e.g., disease improvement metrics, control group, baseline measurements). Since causal claims require comparator outcomes, the effectiveness portion remains unsubstantiated. While tolerability is entailed, the overall label is neutral."*
>
> **CoT (abridged).** *"The statement is supported since only mild AEs were reported . . . this indicates the drug was effective and well tolerated."*

The agentic solver follows the expected *comparator requirement*, decomposing causal claims into subcomponents and assigning neutral when one is unsupported. The CoT baseline collapses tolerability and efficacy into a single plausibility heuristic, producing over-attribution. This demonstrates how structural prompts prevent the conflation of post-treatment outcomes with causal effect. Under correct scaffolding, LLMs reliably enforce comparator rules; without it, they default to association heuristics.

Compositional Grounding (Problem 12).

> **Premise:** *"Fludarabine 120 mg/m^2 daily $\times 14$ days for CLL in an elderly patient."*
> **Statement:** *"The treatment is expected to induce remission, improve blood counts, and prolong survival."*
>
> **Agentic (verbatim).** *"1) Extract factors: fludarabine 120 mg/m^2 daily $\times 14$. 2) Standard dose is 25 mg/m^2/day $\times 5$. The given regimen is $\approx 5\times$ higher and extended. 3) Such dosing causes severe toxicity in elderly patients. 4) This violates therapeutic admissibility; the statement's claim of benefit is contradicted."*
>
> **CoT (abridged).** *"Fludarabine is used in CLL, so this treatment is likely to be effective . . . the statement is entailed."*

The agentic solver operationalises the *cross-factor constraint*, checking consistency across drug, dose, schedule, and patient. This reproduces the knowledge framework's tuple-based admissibility rule. The CoT reasoning reflects the failure mode of Jullien et al. (2025): collapsing multi-factor structures into a single drug–indication match. Here, agentic scaffolding ensures toxicity is identified, whereas CoT ignores unsafe interactions. The repercussion is clinical: a toxic regimen is falsely labelled as beneficial under free CoT, while the structured agent enforces compositional validity.

Epistemic Verification (Problem 16).

> **Premise:** *"A 45-year-old man complains of chest discomfort after meals and occasional regurgitation. Endoscopy is normal. Despite the absence of cardiac symptoms, the physician diagnoses myocardial infarction and starts anticoagulation therapy."*
> **Statement:** *"The patient has myocardial infarction."*
>
> **Agentic (verbatim).** *"1) Endoscopy normal (objective evidence). 2) Symptoms consistent with GERD, not MI. 3) MI requires ECG/troponins, absent here. 4) Physician's diagnosis conflicts with higher-tier evidence. 5) Conclusion: contradiction."*
>
> **CoT (abridged).** *"The physician diagnosed MI, so this supports the statement . . . although there is no cardiac evidence, the statement is entailed."*

The agentic solver follows the *evidence hierarchy*, prioritising instrumented evidence over unsupported assertion. The CoT baseline defers to physician authority, ignoring hierarchy. This misalignment mirrors the epistemic collapse observed previously: models recognise evidence but misapply credibility weights. By enforcing explicit tiering, the agentic pipeline compels the correct contradiction label. This illustrates that LLMs are capable of stratified verification if prompted to discard lower-tier claims in conflict with higher-tier evidence.

Risk State Abstraction (Problem 39).

> **Premise:** *"A 55-year-old man with acute severe low back pain reports saddle anesthesia, urinary retention, and bilateral leg weakness. Reflexes reduced. No imaging performed."*
> **Statement:** *"Emergency MRI is required to exclude cauda equina syndrome."*
>
> **Agentic (verbatim).** *"Symptoms constitute red flags for CES. Likelihood: high, since multiple features align. Severity: catastrophic, with risk of paralysis if missed. No imaging means CES remains unexcluded. Therefore urgent MRI is mandated; statement entailed."*
>
> **CoT (abridged).** *"The patient has back pain and weakness . . . MRI could help but it is not certain; label neutral."*

The agentic solver instantiates the *severity–likelihood calculus*: combining probability with consequence to project latent catastrophic risk. The CoT baseline fails to abstract beyond symptom description, treating CES as one of many possible adverse events. This reproduces the frequency heuristic failure identified in prior work. The consequence is profound: under CoT, a time-critical emergency is missed. The agentic framework ensures catastrophic risk is correctly prioritised, demonstrating that LLMs can perform principled risk abstraction when explicitly instructed.

Across reasoning families, the agentic framework compels alignment with formal inferential schemas: comparators for causality, cross-factor constraints for composition, evidence hierarchies for epistemic verification, and severity–likelihood calculus for risk. The CoT baseline fails in family-specific, systematic ways: collapsing causal claims, flattening compositional structure, deferring to authority, and ignoring catastrophic risk. These are not random lapses but consistent heuristics. The crucial finding is that LLMs *can* reproduce the expected reasoning patterns under correct prompting: the agentic pipeline demonstrates that comparator tests, constraint enforcement, tiered evidence, and risk abstraction are all within model capacity, provided the reasoning path is structured. Without such scaffolding, CoT defaults to surface plausibility, with direct repercussions for clinical safety and validity.

5 Future Work

The principal bottleneck of CARENLI lies in reasoning-family classification, as task success depends critically on accurate routing of items to their appropriate inferential schema. One promising direction is to improve the formalisation of the classification task itself, for example by designing explicit symbolic representations of reasoning type signatures to guide planner decisions. Alternatively, a dedicated planner model could be trained on annotated data to provide more reliable routing.

A second direction concerns *multi-family reasoning*. In realistic clinical inference, problems rarely decompose neatly into a single reasoning type. Extending CARENLI with hierarchical planners capable of invoking multiple solvers sequentially or in parallel would more closely approximate this structure, alleviating the bottleneck of forcing a decision between potentially valid reasoning families.

Third, expanding the repertoire of reasoning families offers a natural path toward broader applicability. Beyond the current four families, future extensions could incorporate temporal progression, counterfactual safety, and resource-constrained decision-making. Such additions would enable CARENLI to capture a wider spectrum of clinically relevant inference patterns.

Finally, while the present evaluation is conducted on controlled, template-derived datasets, future work should assess CARENLI in less constrained settings, such as MedNLI (Romanov and Shivade, 2018), and NLI4CT (Jullien et al., 2023), offering a more stringent test of the framework's robustness and clinical utility.

6 Related Work

A prevailing assumption in NLP is that enlarging model capacity and exposure will improve not only task performance but also the fidelity of underlying reasoning. Scaling law analyses (Kaplan et al., 2020; Hoffmann et al., 2022) and flagship model reports (Brown et al., 2020; Touvron et al., 2023; Bubeck et al., 2023) reinforce this view, yet critics observe that benchmark success often reflects surface regularities rather than robust inference (Marcus, 2022; Mahowald et al., 2024). Our work instantiates this critique in the clinical domain: we show that models encode relevant medical facts but fail to deploy them within principled reasoning flows.

Evidence across NLI research supports this diagnosis. General-domain studies reveal annotation artifacts and shortcut reliance (Gururangan et al., 2018; Poliak et al., 2018; Webson and Pavlick, 2022; Turpin et al., 2023). Clinical NLI benchmarks extend these concerns: MedNLI (Romanov and Shivade, 2018), NLI4CT (Jullien et al., 2023), and CTNLI (Jullien et al., 2025) report systematic reasoning errors, motivating frameworks that explicitly separate knowledge retrieval from inferential schema adherence.

Parallel to this diagnostic agenda, recent work has investigated agentic and modular architectures as a pathway to more reliable reasoning. Generative Agents (Park et al., 2023) and multi-agent frameworks such as CAMEL (Li et al., 2023), AutoGen (Wu et al., 2024), and MetaGPT (Hong et al., 2024) show that distributing tasks across specialised roles facilitates self-correction and auditability. CARENLI extends this line by introducing compartmentalisation grounded in formally defined reasoning families, combining the auditability of agentic design with the domain fidelity demanded by clinical inference.

7 Conclusion

We introduced *CARENLI*, a Compartmentalised Agentic Reasoning framework for Clinical NLI that decomposes inference into planner, solver, verifier, and refiner roles aligned with four reasoning families. By enforcing modular separation and grounding each component in domain-specific procedures, CARENLI enables principled alignment and transparent error diagnosis.

Empirical evaluation across four contemporary LLMs demonstrated substantial gains in reasoning fidelity relative to agnostic prompting, with improvements of up to 42 points and near-ceiling performance in tasks such as causal attribution and compositional grounding under oracle planning. The inclusion of verifier and refinement stages further enhanced reliability: reaching perfect accuracy in risk abstraction, while refinement corrected up to 39% of errors in epistemic tasks by enforcing factual and structural consistency.

These findings indicate that structured agentic prompting provides a viable pathway for aligning LLM reasoning with domain-grounded schemas, supporting safe and auditable deployment in clinical decision-making

References

2017. Common terminology criteria for adverse events (ctcae) v5.0. `https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm`.

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, and Khalid H Malki. 2024. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, 12(1):e54345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, Stephen Arbuck, Steve Gwyther, Margaret Mooney, et al. 2009. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. 2004. *Reasoning about knowledge*. MIT press.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

MG Myriam Hunink, Milton C Weinstein, Eve Wittenberg, Michael F Drummond, Joseph S Pliskin, John B Wong, and Paul P Glasziou. 2014. *Decision making in health and medicine: integrating evidence and values*. Cambridge university press.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press.

Maël Jullien, Marco Valentino, and André Freitas. 2025. The knowledge-reasoning dissociation: Fundamental limitations of llms in clinical natural language inference. *arXiv preprint arXiv:2508.10777*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023. Nli4ct: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

David Lewis. 1973. Causation. *The journal of philosophy*, 70(17):556–567.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.

Gary Marcus. 2022. Deep learning is hitting a wall. *Nautilus*, 10:2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Xin Quan, Marco Valentino, Louise A Dennis, and Andre Freitas. 2024. Enhancing ethical explanations of large language models through iterative symbolic refinement. *arXiv preprint arXiv:2402.00745*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Steven Sloman and Steven A Sloman. 2009. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Johan Van Benthem. 2011. *Logical dynamics of information and interaction*. Cambridge University Press.

Bas C Van Fraassen. 1980. *The scientific image*. Oxford University Press.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2300–2344.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

A Appendix

	Causal Attr.		**Comp. G.**		**Epis. Verif.**		**Risk Abstr.**	
Model	Δ	Gain	Δ	Gain	Δ	Gain	Δ	Gain
CARENLI								
gemini	0.0	–	0.0	–	1.3	50.0	0.7	100.0
deepseek-r1	2.0	100.0	8.0	16.7	8.7	73.3	5.4	0.0
gpt-4o	6.0	100.0	6.2	66.7	4.7	100.0	6.1	60.4
gpt-4o-mini	28.0	76.7	22.2	23.3	39.3	54.2	16.6	60.0
Oracle Planner								
gemini	0.0	–	0.0	–	0.0	–	5.0	0.0
deepseek-r1	0.0	–	5.0	0.0	5.0	0.0	0.0	–
gpt-4o	10.0	100.0	15.0	100.0	5.0	0.0	0.0	–
gpt-4o-mini	25.0	100.0	35.0	28.6	30.0	0.0	0.0	–

Table 5: Mind-change summary across reasoning tasks, showing only change rate (Δ) and proportions of gains.