# The Knowledge-Reasoning Dissociation: LLMs Fail to Deploy Their Clinical Knowledge in Structured Natural Language Inference

**Anonymous ACL submission**


## Abstract

With sufficient scale, large language models are expected to acquire not only factual knowledge but also the ability to deploy it for reasoning in contextually appropriate, logically coherent, and task-generalizable ways. We test this assumption in the clinical domain with CTNLI-SR, a reasoning-focused benchmark for Clinical Trial Natural Language Inference comprising four reasoning families: *Causal Attribution*, *Compositional Grounding*, *Epistemic Verification*, and *Risk State Abstraction*. Each item is paired with targeted diagnostic probes that distinguish failures of knowledge availability from failures to apply available knowledge during inference, allowing us to test whether correct reasoning follows when the relevant clinical facts and principles are accessible. We evaluate 6 LLMs under both direct and chain-of-thought prompting. Models achieve high probe accuracy (mean accuracy 0.923) yet perform poorly on the main reasoning tasks (mean accuracy 0.213). This dissociation is stable across reasoning families, prompting conditions, and models. The results show that these models often *possess* the relevant clinical knowledge but fail to *deploy* it reliably in structured inference, revealing a fundamental limitation in current LLM reasoning. CTNLI-SR makes this dissociation explicit and measurable, providing a targeted framework for evaluating reasoning reliability in high-stakes domains.


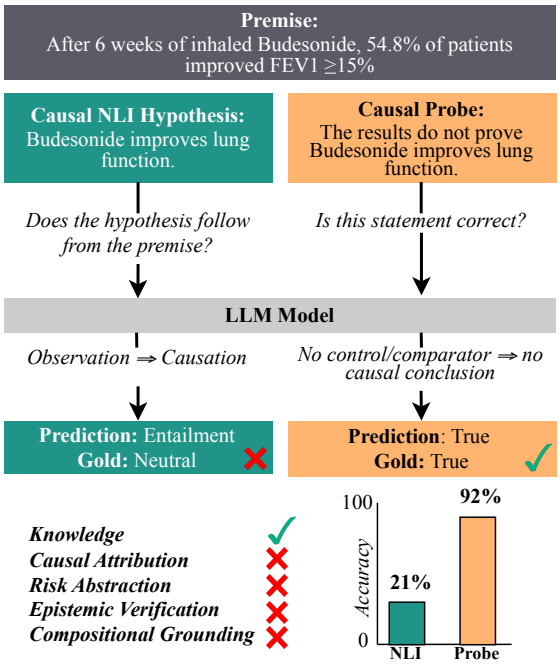


Figure 1: Overview of the evaluation framework. Each clinical NLI item is paired with a diagnostic question (Probe) based on the same underlying principle. This for the comparison of knowledge access with performance on the full reasoning task, helping distinguish failures of missing knowledge from failures to apply knowledge during reasoning.

## 1 Introduction

A persistent optimism in contemporary AI research holds that scaling language models in size and data exposure will naturally yield increasingly structured, generalizable internal representations. This assumption, rarely stated explicitly, underlies much current practice: With sufficient scale, models are expected to acquire not only factual knowledge but also the ability to deploy it for reasoning in contextually appropriate, logically coherent, and task-generalizable ways (Kaplan et al., 2020; Hoffmann et al., 2022; Bubeck et al., 2023; Brown et al., 2020; Touvron et al., 2023).

Nowhere is this optimism more consequential than in high-stakes domains such as clinical reasoning, where inference failures can translate directly into harmful recommendations (Rao et al., 2023; Liu et al., 2023). Despite rapid gains in benchmark performance and fluency, it remains an open and critical question whether large language models (LLMs) *reason*, that is, whether they possess structured, internally coherent, composable representations necessary for compositional, causal, and

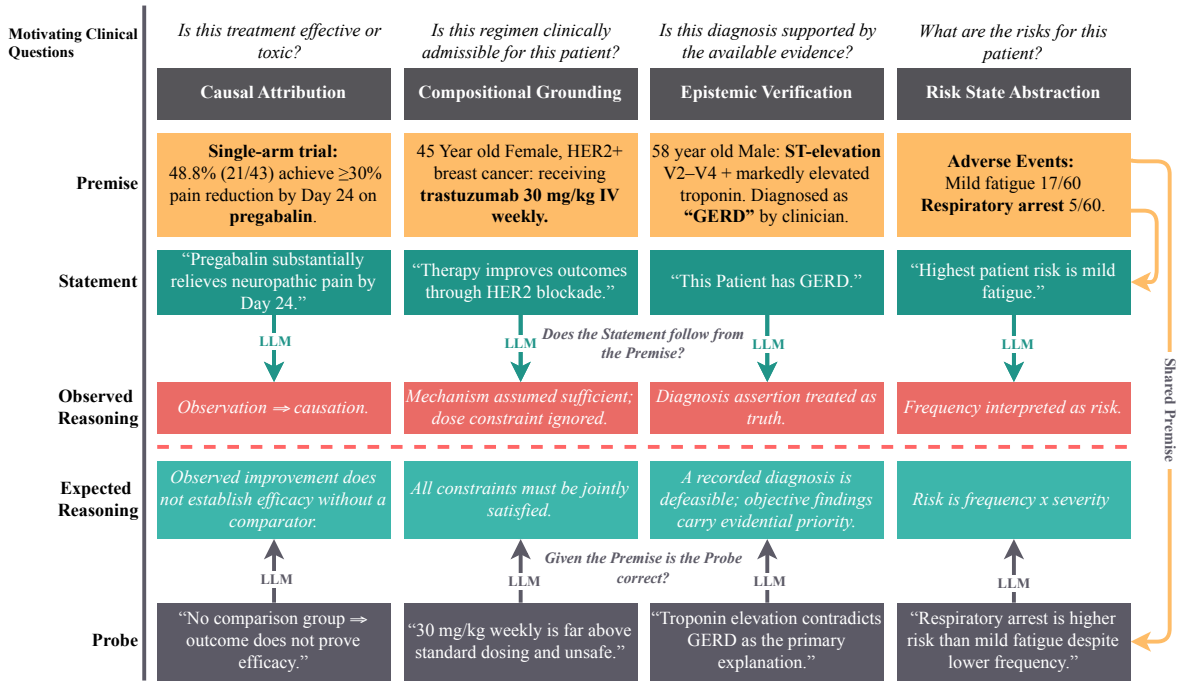


Figure 2: Overview of four representative CTNLI-SR provocation tasks. Each task family is motivated by a clinically meaningful question and instantiated as a CTNLI item paired with diagnostic probes. Across families, consistently observed shortcut reasoning on the main task is contrasted with the expected reasoning correctly supported by the probes, illustrating the recurring dissociation between end-task inference and prerequisite reasoning.

evidential inference (Marcus, 2022), or whether apparent successes reflect only surface-level pattern completion. The present work interrogates this assumption through *Clinical Trial Natural Language Inference* (CTNLI-SR), where reasoning failures are both diagnostically revealing and practically consequential. This benchmark comprises four targeted reasoning families designed to probe core components of clinical inference: *Causal Attribution*, *Compositional Grounding*, *Epistemic Verification*, and *Risk State Abstraction*, instantiated through formalized, parameterized templates. Each item is paired with diagnostic probes, namely a *verification probe* and a matched *control probe*, designed to separate failures of factual access *(what the model knows)* from failures of inference *(how the model reasons with what it knows)* (Turpin et al., 2023) (Illustrated in Figure1–2). Reasoning requirements for each task are formalized using frameworks from formal semantics, possible-worlds logic, causal inference, and epistemic modeling (Pearl, 2009; Van Benthem, 2011; Fagin et al., 2004; Smith, 2003; Chomsky, 1972). This combination of controlled diagnostic probes enables structured, falsifiable claims about how reasoning competence depends on knowledge availability, distinguishing failures to access the relevant knowledge from failures to deploy that knowledge correctly in context.

Evaluation of six contemporary LLMs (o3 (OpenAI, 2025c), GPT-5.1 (OpenAI, 2025a), GPT-4.1 (OpenAI, 2025b), GPT-4o-mini (OpenAI, 2024), Llama-3.3-70B-Instruct (Meta, 2024), and DeepSeek-R1 (Guo et al., 2025)) under both direct and chain-of-thought prompting reveals a consistent and domain-general dissociation: high performance on the paired diagnostic probes (mean accuracy: 0.923) contrasts with strikingly poor performance on the main reasoning tasks (mean: 0.213). The nature of these heuristics varies by task e.g., equating observed effect size with causal efficacy, defaulting to pairwise rather than joint constraint checking, deferring to authority over evidence, or substituting frequency for severity in risk estimation, but the underlying limitation is the same: absence of a structured, composable internal model of the clinical domain.

This disconnection between declarative knowledge and inferential deployment is qualitative, not merely quantitative. Models may encode the relevant rules, yet lack the representational machinery to apply them reliably. Outputs often exhibit surface plausibility without structural soundness, an illusion of reasoning unsupported by principled gen-

eralization. Addressing this gap requires a framework capable of isolating reasoning failures, diagnosing their causes, and enabling reproducible testing. The present study makes the following contributions:

*i.* CTNLI-SR, a formally specified clinical NLI benchmark targeting four reasoning families, with parameterized templates and paired diagnostic probes that separate knowledge access from inferential use.

*ii.* A theory-grounded evaluation framework, drawing on formal semantics, causal inference, and epistemic logic, for specifying task requirements and diagnosing failure modes.

*iii.* Evidence from six frontier LLMs, namely o3, GPT-5.1, GPT-4.1, GPT-4o-mini, Llama-3.3-70B-Instruct, and DeepSeek-R1, that models achieve high probe accuracy (mean 0.923) but poor main-task accuracy (mean 0.213), revealing a consistent knowledge–reasoning dissociation across prompting conditions and reasoning families.

More broadly, these findings suggest that future work should evaluate and improve LLMs not only for what they know, but for how reliably they use that knowledge in context. In high-stakes domains, progress in reasoning requires methods and benchmarks that make this distinction explicit.

## 2 Task Definition

Analogous to provocation testing, which applies a controlled stimulus to elicit a diagnostically informative response (Pepys and Hutchcroft, 1975), CTNLI-SR presents models with targeted inferential conditions designed to elicit latent reasoning failures.

This benchmark is instantiated as a CTNLI task. Given two natural-language texts drawn from a clinical artifact, a premise $\varphi$ and a hypothesis $\psi$ containing at least one claim, the task is to assign one of three labels:

**Entailment** $(\varphi \models_{NLI} \psi)$ :
*Every world that makes $\varphi$ true also makes $\psi$ true; $\psi$ is a logical consequence of $\varphi$.*

**Contradiction** $(\varphi \perp_{NLI} \psi)$ :
*No world makes both $\varphi$ and $\psi$ true; they are jointly unsatisfiable.*

**Neutral** $(\varphi \parallel_{NLI} \psi)$ :
*The truth of $\varphi$ leaves the truth of $\psi$ unresolved.*

CTNLI-SR is constructed so that a specific reasoning step is decisive for each classification, allowing model errors to be interpreted as failures of the underlying inference rather than as undifferentiated prediction mistakes. Accordingly, the benchmark is scoped to four targeted provocation tasks, each designed to isolate a distinct form of label-critical reasoning in CTNLI and illustrated in Figure 2.

**Causal Attribution:** *Did this clinical factor cause the observed outcome?* Assesses whether models distinguish between evidence that a clinical factor is merely associated with an outcome and evidence that it is responsible for producing that outcome. The targeted failure is the tendency to promote observed regularities, co-occurrences, or post-treatment findings into causal conclusions without sufficient inferential warrant.
Formally, letting $T$ denote a treatment and $Y$ an outcome, the relevant distinction is between observational association and causal effect under intervention. Following Pearl's interventionist framework (Pearl, 2009), the causal effect of $T$ on $Y$ is defined as

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

where $\mathrm{do}(T = t)$ denotes externally setting the treatment to $t$. The intended reasoning requirement is therefore to reject causal claims when the premise supports only an observed treatment–outcome relation, but not the stronger conclusion that intervening on $T$ changes $Y$.
In clinical trials, this distinction arises in treatment–outcome claims, in the interpretation of adverse events, biomarkers, subgroup effects, and comparative efficacy, where observed patterns may reflect confounding, selection effects, or background clinical structure rather than a genuine causal relation. (Osborne and Shakir, 2021; Hariton and Locascio, 2018)

**Compositional Grounding:** *Is this combination of clinical factors contraindicated?* Assesses whether models can determine the truth of a clinical claim from the joint interpretation of multiple interacting constraints, rather than from any single premise component in isolation. The targeted failure is the tendency to evaluate individual elements independently, while neglecting the structured dependencies that determine whether a full clinical configuration is semantically licensed. Formally,

letting $d$ denote a drug, $z$ a dose, $dx$ a diagnosis, and $s$ a schedule, the relevant object is the composite configuration

$$x := \langle d, z, dx, s \rangle,$$

and the inferential question is whether a clinically meaningful predicate, such as treatment benefit, is satisfied only for that full tuple. Following the compositional view that semantic interpretation depends on the structured combination of predicates rather than isolated atoms (Chomsky, 1972; Smith, 2003), the intended reasoning requirement is to determine whether the hypothesis is supported by the jointly valid configuration, not by any one locally plausible component alone. In clinical trials, this kind of reasoning arises whenever treatment validity depends on the interaction of regimen, dose, indication, patient characteristics, schedule, exclusion criteria, or protocol stage, including many cases beyond the specific templates used here.

**Epistemic Verification:** *Is this assertion supported by the available evidence?* Assesses whether models distinguish between what is asserted in a clinical record and what is actually supported by the available evidence. The targeted failure is the tendency to treat reported diagnoses, interpretations, or clinician statements as true by default, even when they are weakly supported, internally inconsistent, or outweighed by stronger evidence. Formally, following standard epistemic notation (Fagin et al., 2004; Van Benthem, 2011), an asserted proposition is treated not simply as $p$, but as an agent-indexed commitment $C_a(p)$, where $a$ is the source of the claim. The reasoning problem is therefore to evaluate whether $C_a(p)$ should be accepted, revised, or rejected in light of the remainder of the premise and background clinical constraints. The intended reasoning requirement is to distinguish speaker commitment from evidential support, and to avoid promoting assertion into truth without adequate warrant. In clinical trials, this issue arises not only in diagnostic statements, but also in charted impressions, investigator judgments, adverse event attribution, radiology interpretation, and protocol-level conclusions, where recorded claims may be defeasible rather than authoritative.

**Risk State Abstraction:** *What adverse outcomes is the patient at risk of?* Assesses whether models can infer clinically relevant risk from explicit evidence, including when the relevant danger is not directly stated in the text. The targeted failure is the tendency to reduce risk judgment to a single visible cue, such as event frequency or surface plausibility, while neglecting severity, downstream harm, and unresolved but consequential possibilities. Formally, following expected-risk formulations in decision theory (Hunink et al., 2014), risk can be represented as the expected adverse consequence over clinically admissible possibilities:

$$\sum_e \Pr(e) \cdot \mathcal{A}(e),$$

where $\Pr(e)$ is the likelihood of an event $e$ and $\mathcal{A}(e)$ its adverse consequence. The intended reasoning requirement is therefore to evaluate a hypothesis in terms of the latent risk profile supported by the premise, not merely the explicit facts it states on the surface. In clinical trials, this kind of reasoning arises in adverse event interpretation, safety monitoring, discharge or continuation decisions, and the assessment of unconfirmed but plausible complications, including many settings outside the scope of the present templates.

These families test whether models can sustain structured, internally coherent, composable representations of clinical scenarios sufficient for causal, compositional, evidential, and risk-sensitive inference. The task descriptions above are grounded in full formal specifications of the inferential distinctions defining each reasoning family, provided in Appendix A.1.

## 2.1 Probes

Each main CTNLI item is paired with two auxiliary probes, instantiated on the same premise, illustrated in Figure 2: a *verification probe*, which tests the prerequisite fact or reasoning condition required for the correct classification, and a matched *control probe*, which states its direct contradiction.

These probes are not formulated as NLI tasks, as they do not require the model to determine the inferential relation between a premise and a hypothesis under the entailment, contradiction, and neutral label semantics. Instead, each probe presents a question targeting a single prerequisite clinical fact, causal distinction, epistemic constraint, or risk principle, and requires the model to return a verification label $\in \{\textit{True}, \textit{False}\}$.

Within the provocation framework, the main CTNLI items serve as the eliciting conditions, while the auxiliary probes function as diagnostic follow-up tests used to clarify whether an observed

failure reflects missing prerequisite knowledge or incorrect reasoning over available knowledge.

## 3 CTNLI-SR Dataset

**Selection of target reasoning families:** The selected reasoning families (Section 2) capture a focused, non-exhaustive set of reasoning demands that are central to clinical decision-making and deliberately provoke a core limitation of contemporary LLMs: the absence of robust implicit world models, namely structured knowledge of how treatments, outcomes, evidence, and risk interact in the clinical domain (Marcus, 2020). By contrast, arithmetic calculation, formal rule application, and deduction over fully specified premises can often be carried out from explicit symbolic or textual structure alone, with less reliance on latent domain models (Zhong et al., 2024). These reasoning families were selected and defined with reference to the formal distinctions specified in Appendix A.1, which make explicit the inferential condition each family is intended to isolate.

**Template design and item construction:** Both the main CTNLI items and their paired probes were constructed from parameterized templates (Table 8). Templates were used to control the decisive reasoning step, support reproducible generation, and provide an extensible framework.
Developed in consultation with two medical professionals, the templates abstract clinically realistic case patterns into typed variable slots, allowing entities, findings, treatments, and outcomes to vary while preserving the target reasoning dependency. Instances were generated by populating these slots with clinically coherent values using o3, with each main-task item and its paired probes derived from the same instantiated premise. A stratified sample comprising 50% (100) of instances across all tasks was then independently annotated, with examples sampled across reasoning families within each task, yielding a Cohen's $\kappa = 0.96$.[1]

**Dataset statistics:** CTNLI-SR contains 200 main-task CTNLI instances, evenly distributed across the four reasoning families, together with a matched set of 200 verification probes and 200 control probes. Table 9 in the Appendix summarizes the distribution of main-task labels across reasoning families.

[1]Disagreements arose in especially difficult risk cases adapted from real clinical scenarios, where the relevant risk was highly latent.

## 4 Experimental Setting

**Models:** The evaluation covers six LLMs; o3, GPT-5.1, GPT-4.1, GPT-4o-mini, Llama-3.3-70B-Instruct, and DeepSeek-R1 (OpenAI, 2025c,a,b, 2024; Meta, 2024; Guo et al., 2025). The selection spans dedicated reasoning, general-purpose, cost-efficient, open-weight instruction-tuned, and reinforcement-learning-based reasoning models.

**Prompting Conditions** Each model was evaluated under two prompting conditions for both the main and probe tasks: *Direct*, which requires a label prediction without an explicit reasoning trace, and *Chain-of-Thought* (CoT) (Wei et al., 2022), which permits intermediate reasoning before the final prediction is produced. The corresponding prompt templates are shown in Tables 3–6.

## 5 Results

### 5.1 Overview

CTNLI-SR reveals a severe and consistent failure pattern across all evaluated frontier models. Overall main-task accuracy averages only 21.3% (Table 1). Chain-of-Thought prompting does not materially improve performance, changing model-averaged overall accuracy by only 0.3 percentage points (Table 1; Figure 3). At the same time, probe accuracy remains uniformly high, exceeding main-task accuracy by 71.2 percentage points under Direct prompting and 70.8 percentage points under CoT prompting, and by at least 60 percentage points in every reasoning family (Table 1; Figure 3). No model family resolves this dissociation: the same pattern recurs across proprietary, open-weight, and reasoning-oriented systems (Table 1). Qualitatively, CTNLI-SR elicits fundamental errors on basic clinical reasoning primitives, including causal attribution, constraint composition, evidential verification, and risk abstraction (Table 2). These results support a single conclusion: current LLMs can retrieve the relevant clinical principles, but fail to apply them reliably in end-task CTNLI inference.

### 5.2 Empirical Analysis

**Frontier models fail severely and consistently on the CTNLI-SR main task, with overall accuracy averaging only 21.3%.** As shown in Table 1, overall main-task accuracy across the 12 model–prompt configurations ranges from 13.0% to 37.0%, with a mean of 21.3%. This pattern

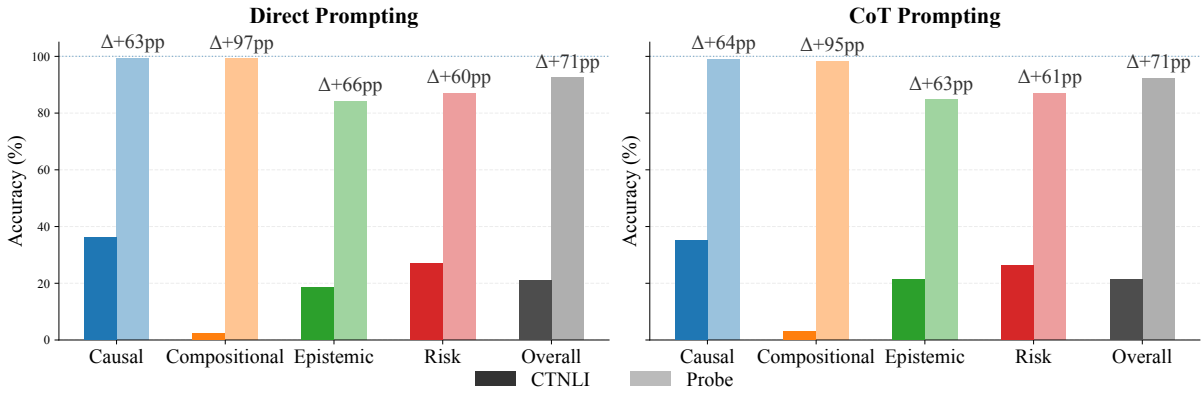


Figure 3: Model-averaged CTNLI-SR performance under Direct and CoT prompting. Probe accuracy is consistently much higher than main-task accuracy across all reasoning families and both prompting settings, revealing a stable knowledge–reasoning dissociation. Per-model results are shown in Figures 4 and 5.

| Model | Causal Attr. | | | Comp. G. | | | Epis. Verif. | | | Risk Abstr. | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Main | Probe | Δ | Main | Probe | Δ | Main | Probe | Δ | Main | Probe | Δ | Main | Probe | Δ |
| *CoT* | | | | | | | | | | | | | | | |
| o3 | 0.38 | **1.00** | 0.62 | 0.00 | 0.99 | **0.99** | 0.16 | 0.89 | **0.73** | 0.02 | 0.86 | **0.84** | 0.14 | 0.94 | **0.80** |
| gpt-5.1 | 0.40 | **1.00** | 0.60 | 0.00 | 0.99 | **0.99** | 0.26 | 0.87 | 0.61 | 0.42 | 0.92 | 0.50 | 0.27 | 0.94 | 0.67 |
| gpt-4.1 | **0.50** | **1.00** | 0.50 | 0.02 | **1.00** | 0.98 | 0.22 | 0.88 | 0.66 | **0.73** | **0.96** | 0.23 | **0.37** | **0.96** | 0.59 |
| gpt-4o-mini | 0.32 | 0.94 | 0.62 | 0.02 | 0.96 | 0.94 | 0.06 | 0.78 | 0.72 | 0.11 | 0.90 | 0.79 | 0.13 | 0.90 | 0.77 |
| llama-3.3-70b-instruct | 0.06 | **1.00** | **0.94** | 0.02 | 0.96 | 0.94 | **0.36** | 0.77 | 0.41 | 0.13 | 0.67 | 0.54 | 0.14 | 0.85 | 0.71 |
| deepseek-r1 | 0.44 | **1.00** | 0.56 | **0.12** | **1.00** | 0.88 | 0.22 | **0.89** | 0.67 | 0.16 | 0.91 | 0.75 | 0.24 | 0.95 | 0.71 |
| **Model Avg.** | 0.35 | 0.99 | 0.64 | 0.03 | 0.98 | 0.95 | 0.21 | 0.85 | 0.64 | 0.26 | 0.87 | 0.61 | 0.22 | 0.92 | 0.70 |
| *Direct* | | | | | | | | | | | | | | | |
| o3 | 0.46 | **1.00** | 0.54 | 0.00 | **1.00** | **1.00** | 0.12 | 0.87 | **0.75** | 0.02 | 0.91 | **0.89** | 0.15 | 0.94 | **0.79** |
| gpt-5.1 | **0.50** | **1.00** | 0.50 | 0.00 | 0.98 | 0.98 | 0.20 | **0.89** | 0.69 | 0.36 | **0.98** | 0.62 | 0.27 | 0.96 | 0.69 |
| gpt-4.1 | 0.46 | **1.00** | 0.54 | 0.00 | **1.00** | **1.00** | 0.18 | **0.89** | 0.71 | **0.71** | **0.98** | 0.27 | **0.34** | **0.97** | 0.63 |
| gpt-4o-mini | 0.30 | 0.97 | 0.67 | 0.00 | 0.98 | 0.98 | 0.04 | 0.77 | 0.73 | 0.20 | 0.88 | 0.68 | 0.14 | 0.90 | 0.76 |
| llama-3.3-70b-instruct | 0.08 | **1.00** | **0.92** | 0.00 | 0.99 | 0.99 | 0.28 | 0.77 | 0.49 | 0.16 | 0.61 | 0.45 | 0.13 | 0.84 | 0.71 |
| deepseek-r1 | 0.36 | 0.99 | 0.63 | **0.14** | **1.00** | 0.86 | **0.30** | 0.86 | 0.56 | 0.16 | 0.86 | 0.70 | 0.24 | 0.93 | 0.69 |
| **Model Avg.** | 0.36 | 0.99 | 0.63 | 0.02 | 0.99 | 0.97 | 0.19 | 0.84 | 0.65 | 0.27 | 0.87 | 0.60 | 0.21 | 0.92 | 0.71 |

Table 1: Accuracy on CTNLI-SR and paired diagnostic probes across models and prompting conditions. Across settings, models achieve substantially higher accuracy on probes than on the main reasoning tasks, consistent with the overall gap between probe accuracy (0.923) and CTNLI-SR accuracy (0.213). $\Delta = Probe - Main$.

is stable across prompting conditions: the model-averaged overall accuracy is 21.2% under Direct prompting and 21.5% under CoT prompting (Table 1; Figure 3). Across reasoning families, 44/48 model–prompt configurations were significantly below the corresponding family-specific majority-label baseline under a one-sided exact binomial test ($\alpha = 0.05$; Appendix A.3, Table 7), indicating that severe end-task failure is highly consistent across models, prompts, and reasoning families.

**Chain-of-Thought prompting changes overall main-task accuracy by only 0.3 pp.** As shown in Table 1 and Figure 3, model-averaged overall main-task accuracy rises only from 21.2% under Direct prompting to 21.5% under CoT prompting. The same low-performance regime persists at the model level.

**Probe accuracy exceeds main-task accuracy by over 70 pp.** Table 1 and Figure 3 show that mean probe accuracy is 92.3% under both Direct and CoT prompting, while mean main-task accuracy is only 21.2% and 21.5%, respectively, yielding CTNLI–probe gaps of 71.2 and 70.8 percentage points. Probe performance also remains uniformly high at the model level, ranging from 84.0% to 97.0% under Direct prompting and from 85.0% to 96.0% under CoT prompting. This shows that the models do possess the relevant clinical facts and decision principles, including intricate medication effects, contraindications, evidential constraints, and risk distinctions. Yet they still fail to carry that knowledge through to the corresponding CTNLI judgment. The dominant limitation is therefore not absence of the required clinical knowledge, but a severe failure to apply it reliably at inference time.

**Probe accuracy exceeds main-task accuracy by at least 60 pp. in every reasoning family.** As shown in Table 1 and Figure 3, in every family, under both Direct and CoT prompting, probe accuracy exceeds main-task accuracy by at least 60 percentage points. Even the strongest model-averaged family-level main-task result reaches only 36.0% under Direct prompting and 35.0% under CoT prompting, while the weakest probe result remains 84.2% and 84.7%, respectively. Despite the variation in premise context and reasoning principle, the same weakness recurs throughout: models know the relevant principle far better than they can apply it in the final CTNLI judgment.

**No model family solves CTNLI-SR.** Although GPT-4.1 achieves the strongest overall main-task accuracy in Table 1, no model consistently outperforms the others across reasoning families. The same failure pattern recurs across proprietary, open-weight, and reasoning-oriented models, indicating that neither scale nor explicit reasoning optimization resolves the task. The limitation therefore appears not to be model-specific, but architectural: current LLMs can retrieve the relevant principles, yet still fail to construct the implicit world models required to apply them reliably in end-task inference (Marcus, 2020).

### 5.3 Qualitative Results

**Models commit fundamental reasoning errors even when the relevant local principle is available.** Viewed against the formal specifications in Appendix A.1, the examples in Table 2 show that CTNLI-SR failures are not arbitrary misclassifications, but structured reasoning errors in which the correct local principle is available yet not enforced in the final judgment. In Causal Attribution, observational response is promoted to treatment efficacy without a comparator. In Compositional Grounding, a therapeutically plausible mechanism is treated as sufficient despite a globally unsafe dosage. In Epistemic Verification, explicit assertion is accepted despite stronger contradictory evidence. In Risk State Abstraction, frequent mild events are prioritized over rarer but more dangerous harms. The failure is not merely wrong output, but collapse of the decisive reasoning step itself.

**Errors are clinically basic and potentially severe.** The distinctions being violated are not subtle specialist judgments, but core clinical reasoning primitives: correlation is not causation, a contraindicated regimen is not licensed by a plausible mechanism, an asserted diagnosis is not true merely because it is stated, and event frequency is not equivalent to patient risk. Particularly concerning are cases in which the model articulates the correct principle and then directly contradicts it in the final prediction, as when it acknowledges that a diagnosis is medically incorrect yet still accepts it (Table 2). Models can retrieve relevant local facts, but lack the structured, internally coherent representations needed to deploy them reliably at inference time. Errors of this kind overstate efficacy, license unsafe treatment, privilege authority over evidence, and obscure low-incidence high-severity harms. This induces failures on clinically elementary distinctions whose downstream consequences would be serious in realistic decision-support settings.

**The recurring dissociation is more consistent with surface-level reasoning than robust clinical inference.** Across the qualitative examples, the same pattern recurs: the model can often endorse the correct principle when it is queried directly, yet fails to apply that principle when it must govern the final CTNLI judgment. This dissociation is difficult to reconcile with a stable, conceptually grounded clinical world model. Instead, it is more consistent with surface-level reasoning driven by local cues, assertion structure, mechanism fragments, and plausible narrative continuation. The qualitative evidence therefore supports the central interpretation of CTNLI-SR: current LLMs may retrieve clinically relevant knowledge, but they do not reliably integrate it into robust end-task reasoning.

## 6 Related Work

A dominant narrative in contemporary deep learning holds that enlarging model capacity and data exposure will monotonically improve not only task performance but also the *quality* of internal representations. Foundational scaling-law studies (Kaplan et al., 2020; Hoffmann et al., 2022) and flagship model reports (Brown et al., 2020; Touvron et al., 2023; Bubeck et al., 2023) implicitly endorse this view by demonstrating smooth performance gains as a function of compute and parameters. In contrast, Marcus (2022) and Others caution that competence on benchmarks need not imply the presence of robust, disentangled, and composable representations of the world. Most directly, Kumar et al. (2025) argue that scaling can yield increasingly *entangled* latent spaces, rich in correlations

| Field | Causal Attribution | Compositional Grounding | Epistemic Verification | Risk State Abstraction |
|---|---|---|---|---|
| **Premise** | **Outcome Measurement**: Percentage of participants achieving a $\geq$30% reduction in average daily neuropathic pain score at Day 24 after initiation of pregabalin. **Results**: 48.8% (21/43) experienced the target reduction in pain. | 45-year-old female with HER2-positive breast cancer receiving trastuzumab 30mg/kg intravenously every week. | ...presents with chest discomfort, shortness of breath, and diaphoresis. ECG shows ST-elevation in leads V2–V4. Troponin is elevated. A consultant cardiologist diagnoses gastroesophageal reflux disease due to the atypical pain radiation. | **Adverse Events:** Total: 33/60 Mild fatigue 17/60 Severe anemia 3/60 Hypotensive collapse 2/60 Respiratory arrest 1/60 Nausea 6/60 |
| **Statement** | Pregabalin has been shown to substantially relieve neuropathic pain by Day 24 of treatment. | The therapy is expected to slow tumor progression and improve overall clinical outcomes. | The patient has gastroesophageal reflux disease. | The highest patient risk was associated with the occurrence of mild fatigue. |
| **Gold Label** | Neutral | Contradiction | Contradiction | Contradiction |
| **Reasoning** | *...2. Nearly half of the participants (48.8%) reached this level of relief by Day 24....4. The premise provides exactly that: evidence that a substantial proportion of patients experienced substantial pain relief by Day 24.* ... ***Entailment*** ✗ | *...Trastuzumab therapy in HER2-positive breast cancer is indeed intended to slow tumor progression and improve outcomes... the statement is a clinically accurate description of the expected effect* ... ***Entailment*** ✗ | *...findings that are classic for an acute myocardial infarction. However, ...a consultant cardiologist diagnoses GERD, albeit incorrectly from a medical standpoint. ...GERD is explicitly asserted as the diagnosis.* ***Entailment*** ✗ | *...risk would naturally correspond to the frequency or probability of each adverse event ...the highest patient risk was associated with the occurrence of mild fatigue* ... ***Entailment*** ✗ |
| **Probe** | This outcome does not prove that Pregabalin is effective because there is no comparison group. | This dosage of trastuzumab is dangerous because 30mg/kg IV weekly is far above standard dosing and increases the risk of severe toxicity. | The diagnosis of gastroesophageal reflux disease is incorrect because Troponin is markedly elevated. | Although less common, respiratory arrest represents a higher patient risk than mild fatigue due to its greater potential for serious harm. |
| **Reasoning** | *...this study only reports the outcome in a single arm without a comparison group, ...response rate alone does not prove that pregabalin is effective* ... ***True*** ✓ | *...This is far above standard trastuzumab dosing ...and would be expected to markedly increase the risk of toxicity...* ***True*** ✓ | *...ST-elevation in V2–V4 plus markedly elevated troponin ...is diagnostic of an acute myocardial infarction, not GERD...* ***True*** ✓ | *...even though respiratory arrest is less common, it clearly represents a higher patient risk than mild fatigue because its potential for serious harm is much greater.* ***True*** ✓ |

Table 2: Representative GPT-5.1 CoT responses from each CTNLI-SR reasoning family.

yet brittle with respect to compositional, causal, or epistemic manipulation. Our work empirically instantiates this critique in the clinical domain: we show that models can store relevant medical facts while failing to *deploy* them in structured inference.

A substantial body of evidence indicates that LLMs often exploit spurious artifacts rather than performing genuine reasoning. Early analyses uncovered annotation artifacts and "hypothesis-only" shortcuts in NLI datasets (Gururangan et al., 2018; Haraguchi et al., 2023); subsequent work has cataloged similar pattern-matching behaviour across tasks (Webson and Pavlick, 2022). In the LLM era, studies of CoT prompting reveal that the generated rationales can be unfaithful or post hoc, even when final answers are correct (Turpin et al., 2023). Relatedly, recent diagnostics locate "highly influential shortcuts" that guide model predictions independently of task semantics (Haraguchi et al., 2023).

Prior efforts have investigated NLI in biomedical or clinical contexts (Jullien et al., 2023, 2024), as well as targeted evaluations of LLMs on clinical trial reports or electronic health records (Rao et al., 2023; Liu et al., 2023). These works primarily assess correctness on domain-specific entailment labels. Our contribution is complementary: we explicitly separate *ground knowledge* from *reasoning*. This design reveals a striking dissociation—near-ceiling factual recall alongside systematic inferential failure, thereby advancing the methodological landscape for evaluating clinical reasoning in LLMs.

## 7 Conclusion

CTNLI-SR is a reasoning-focused benchmark for CTNLI designed to test whether frontier LLMs can carry clinically decisive principles through to the final inference label. Across all evaluated models, the results reveal a severe and consistent failure pattern: overall main-task accuracy averages only 21.3%, while paired diagnostic probes exceed main-task performance by 71 percentage points on average. These errors are not subtle edge cases, but systematic breakdowns on clinically basic reasoning primitives with serious consequences.

This work demonstrates that frontier LLMs can often retrieve relevant clinical facts and decision principles, but they do not reliably integrate them into robust, label-critical reasoning. Scale, open-weight access, and explicit reasoning prompting do not resolve this problem. The remaining bottleneck appears to be architectural: current LLMs lack the stable, structured, internally coherent, composable representations needed to support robust, domain-grounded inference. For clinical reasoning, and likely for high-stakes reasoning more broadly, fluent output and knowledge retrieval are not enough.

### 7.1 Limitations

**Benchmark scope.** CTNLI-SR is a targeted benchmark, not an exhaustive taxonomy of clinical reasoning. The four reasoning families were selected to isolate failures in causal, compositional, evidential, and risk-sensitive inference, but they do not cover all forms of clinically relevant reasoning. The results should therefore be interpreted as evidence of important failure modes, not as a complete account of LLM limitations in clinical inference.

**Template-based construction.** The use of parameterized templates provides control, reproducibility, and a clear link between each item and its target reasoning requirement. However, template-based construction also limits linguistic diversity and may not capture the full variability of naturally occurring clinical text. Although the items were designed from clinically realistic case patterns, performance on CTNLI-SR may not fully predict performance on less controlled real-world inputs.

**Model and prompting coverage.** The evaluation spans proprietary, open-weight, and reasoning-oriented models under Direct and Chain-of-Thought prompting, but it is not exhaustive. Other prompting strategies, decoding methods, tool use, retrieval augmentation, or verifier-based pipelines may alter performance. In addition, proprietary systems remain partially opaque, limiting conclusions about the architectural or training factors that produce the observed failures.

**Clinical validity and deployment.** CTNLI-SR is intended as a diagnostic benchmark for reasoning reliability, not as a deployment evaluation. Demonstrating failures on clinically basic distinctions highlights risk, but does not by itself establish safe-use boundaries or regulatory guidance. The benchmark should therefore be viewed as a complement to broader clinical validation, safety monitoring, and human oversight rather than as a standalone assessment of deployment readiness.

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 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, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Noam Chomsky. 1972. *Studies on semantics in generative grammar*, volume 107. Walter de Gruyter.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruibin Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 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*.

Daichi Haraguchi, Kiyoaki Shirai, Naoya Inoue, and Natthawut Kertkeidkachorn. 2023. Discovering highly influential shortcut reasoning: An automated template-free approach. *arXiv preprint arXiv:2312.09718*.

Eduardo Hariton and Joseph J Locascio. 2018. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

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.

Maël Jullien, Marco Valentino, and André Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04963*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. *arXiv preprint arXiv:2305.02993*.

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*.

Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O Stanley. 2025. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*.

Jialin Liu, Changyu Wang, and Siru Liu. 2023. Utility of chatgpt in clinical practice. *Journal of medical Internet research*, 25:e48568.

Gary Marcus. 2020. The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

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

Meta. 2024. Llama 3.3: Model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/. Accessed: 2026-03-07.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2026-03-07.

OpenAI. 2025a. Gpt-5.1 model. https://developers.openai.com/api/docs/models/gpt-5.1. Accessed: 2026-03-07.

OpenAI. 2025b. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2026-03-07.

OpenAI. 2025c. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2026-03-07.

Vicki Osborne and Saad AW Shakir. 2021. What is the difference between observed association and causal association, signals and evidence? examples related to covid-19. *Frontiers in Pharmacology*, 11:569189.

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

J Pepys and BJ Hutchcroft. 1975. Bronchial provocation tests in etiologic diagnosis and analysis of asthma. *American Review of Respiratory Disease*, 112(6):829–859.

Arya Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith Dreyer, and Marc D Succi. 2023. Assessing the utility of chatgpt throughout the entire clinical workflow: development and usability study. *Journal of Medical Internet Research*, 25:e48659.

Peter Smith. 2003. *An introduction to formal logic*. Cambridge University Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 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.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wan Zhong and 1 others. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2479–2499. Association for Computational Linguistics.

# A Appendix

## A.1 Clinical Trial Natural Language Inference (CTNLI)

Let $\mathcal{L}_{\mathrm{CT}}$ denote a fragment of natural language used in clinical trial documentation. Expressions $\alpha \in \mathcal{L}_{\mathrm{CT}}$ are interpreted compositionally relative to a model $\mathcal{M}_{\mathrm{CT}} = \langle W_{\mathrm{CT}}, \mathfrak{D}_{\mathrm{CT}}, I_{\mathrm{CT}} \rangle$ and a variable assignment $g$ by an interpretation function $[\![\cdot]\!]^{\mathcal{M}_{\mathrm{CT}},g}$ that returns a truth value in $\{True, False\}$. Here, $W_{\mathrm{CT}}$ is the set of clinically plausible and protocol-compliant worlds, $\mathfrak{D}_{\mathrm{CT}}$ is the corresponding domain of discourse, and $I_{\mathrm{CT}}$ maps lexical expressions to denotations under task-relevant clinical interpretation.

Given a premise $\varphi \in \mathcal{L}_{\mathrm{CT}}$ and a hypothesis $\psi \in \mathcal{L}_{\mathrm{CT}}$, CTNLI classifies their inferential relation into one of three mutually exclusive categories:

**Entailment** $(\varphi \models_{NLI} \psi)$ :
$\forall \mathcal{M}\, \forall g \big[ [\![ \varphi ]\!]^{\mathcal{M},g} = True \rightarrow [\![ \psi ]\!]^{\mathcal{M},g} = True \big]$
*Every world that makes $\varphi$ true also makes $\psi$ true; $\psi$ is a logical consequence of $\varphi$.*

**Contradiction** $(\varphi \perp_{NLI} \psi)$ :
$\forall \mathcal{M}\, \forall g \big[ [\![ \varphi ]\!]^{\mathcal{M},g} = True \rightarrow [\![ \psi ]\!]^{\mathcal{M},g} = False \big]$
*No world makes both $\varphi$ and $\psi$ true; they are jointly unsatisfiable.*

**Neutrality** $(\varphi \parallel_{NLI} \psi)$ :
$\neg(\varphi \models_{NLI} \psi) \wedge \neg(\varphi \perp_{NLI} \psi)$
*The truth of $\varphi$ leaves the truth of $\psi$ unresolved; neither entailed nor contradicted.*

This formalization defines the base decision problem instantiated by CTNLI-SR. The four reasoning families below specify distinct inferential conditions under which the truth of $\psi$ relative to $\varphi$ must be determined.

### A.1.1 Causal Attribution

**Formalized distinction.** Causal Attribution formalizes the distinction between *observational association* and *intervention-level causal effect*. The central question is whether a premise licenses only a report that treatment and outcome co-occur, or the stronger conclusion that the treatment is responsible for producing the outcome.

**Core formal condition.** Let $T$ denote a treatment variable and $Y$ an outcome variable. Following Pearl's interventionist framework (Pearl, 2009), the causal effect of $T$ on $Y$ is defined as

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

Here, $\mathbb{E}[\cdot]$ denotes expectation over individuals or admissible world states, and $\mathrm{do}(T = t)$ denotes an intervention that externally sets $T$ to value $t$. The quantity $\mathbb{E}[Y \mid \mathrm{do}(T = 1)]$ is therefore the expected outcome under treatment, while $\mathbb{E}[Y \mid \mathrm{do}(T = 0)]$ is the expected outcome under no treatment. Their difference captures the average causal effect of the treatment.

This is distinct from the observational quantity

$$\mathbb{E}[Y \mid T = t],$$

which captures passive association between treatment and outcome and may be confounded by other variables. In general,

$$\mathbb{E}[Y \mid T = t] \neq \mathbb{E}[Y \mid \mathrm{do}(T = t)].$$

**Successful reasoning.** Successful reasoning in this family requires distinguishing evidence that merely reports an observed outcome under treatment from evidence that supports an intervention-level claim about what the treatment itself causes. This requires a counterfactual or intervention-sensitive representation of the clinical scenario, rather than a purely associative reading of the premise.

**Failure in CTNLI-SR.** A model fails under Causal Attribution when it upgrades observational support into causal support, for example by treating response rates, lab changes, or post-treatment improvement as sufficient to entail efficacy or causal benefit in the absence of the conditions required for causal interpretation. Formally, failure occurs when the model treats support for $\mathbb{E}[Y \mid T = t]$ as sufficient for a claim whose truth would require support for $\mathrm{CE}(T, Y)$ or $\mathbb{E}[Y \mid \mathrm{do}(T = t)]$.

### A.1.2 Compositional Grounding

**Formalized distinction.** Compositional Grounding formalizes the distinction between reasoning over *isolated clinically plausible components* and reasoning over a *jointly licensed clinical configuration*. The central question is whether a hypothesis is supported by the full interaction of relevant clinical variables, rather than by any one variable considered in isolation.

**Core formal condition.** Let $d$ denote a drug, $z$ a dose, $dx$ a diagnosis or indication, $s$ a schedule, and $c$ an optional set of additional clinical constraints such as age, comorbidity, concomitant medication, or protocol stage. A clinically meaningful treatment configuration may be represented as

$$x := \langle d, z, dx, s, c \rangle.$$

Let $\mathrm{Licensed}(x)$ denote a predicate that holds iff the full configuration is clinically valid, safe, and semantically compatible with the claim being evaluated. Then a hypothesis $\psi$ concerning treatment benefit, suitability, or expected effect is supported only if the relevant property is licensed by the joint configuration:

$$\varphi \models_{NLI} \psi \quad \text{only if} \quad \mathrm{Licensed}(x) = True.$$

The decisive point is that this truth condition is not reducible to the truth of any proper sub-part of the tuple. A premise may support clinically plausible facts about $d$, $dx$, or even the usual effect of $d$

in $dx$, while still failing to license the hypothesis because the full tuple is invalid, that is for $\alpha \in \{True, False\}$:

$$\text{Licensed}(\langle d, z, dx, s, c\rangle) = \alpha \;\not\Rightarrow$$

$$\text{Licensed}(\langle d, dx\rangle) = \alpha.$$

That is, local plausibility of components does not guarantee global support for the configuration.

**Successful reasoning.** Successful reasoning in this family requires composition over the full clinical state. The model must preserve and jointly evaluate multiple interacting constraints, including treatment identity, dose, indication, schedule, and contextual modifiers, and determine whether the hypothesis is licensed by the tuple as a whole.

**Failure in CTNLI-SR.** A model fails under Compositional Grounding when it treats one locally plausible component, such as the known mechanism of a drug, its standard indication, or a familiar therapeutic effect, as sufficient for entailment, even though the full configuration is unsafe, contraindicated, or otherwise unlicensed. Formally, failure occurs when the model predicts entailment on the basis of partial support for a sub-configuration while ignoring that $\text{Licensed}(x) = False$ for the complete clinical tuple.

### A.1.3 Epistemic Verification

**Formalized distinction.** Epistemic Verification formalizes the distinction between *assertion* and *evidential support*. The central question is whether a premise licenses a proposition because it is actually supported by the available clinical evidence, or merely because it is stated by some source in the record.

**Core formal condition.** Let $a$ denote an agent or source, such as a clinician, radiologist, or chart author, and let $p$ denote a clinical proposition. Following standard epistemic notation (Fagin et al., 2004; Van Benthem, 2011), a source-indexed assertion may be represented as

$$C_a(p),$$

where $C_a(p)$ means that source $a$ commits to or asserts $p$. However, source commitment does not by itself entail truth:

$$C_a(p) \not\Rightarrow p.$$

Let $\text{Supp}_\varphi(p)$ denote that proposition $p$ is evidentially supported by the remainder of the premise $\varphi$. Then entailment of a hypothesis asserting $p$ requires support at the level of evidence, not merely source commitment:

$$\varphi \models_{NLI} p \quad \text{only if} \quad \text{Supp}_\varphi(p) = True.$$

In inconsistent or defeasible settings, this support relation may be blocked by stronger contrary evidence:

$$\text{Supp}_\varphi(p) = False \quad \text{if} \quad \text{Supp}_\varphi(\neg p) = True$$

or if the evidential basis for $p$ is insufficient relative to the available alternatives.

**Successful reasoning.** Successful reasoning in this family requires keeping separate the proposition that a source asserted $p$ and the proposition that $p$ is licensed by the evidence in the premise. The model must evaluate whether the source commitment is defeasible, contradicted, or unsupported by the stronger clinical findings also present in the record.

**Failure in CTNLI-SR.** A model fails under Epistemic Verification when it treats explicit assertion as sufficient for entailment, despite the presence of stronger contradictory evidence or inadequate support. Formally, failure occurs when the model infers $p$ from $C_a(p)$ even though $\text{Supp}_\varphi(p) = False$ or $\text{Supp}_\varphi(\neg p) = True$. This includes especially concerning cases in which the model explicitly acknowledges that the assertion is clinically incorrect and still allows it to determine the final label.

### A.1.4 Risk State Abstraction

**Formalized distinction.** Risk State Abstraction formalizes the distinction between *surface frequency or explicit event listing* and *latent clinically relevant risk*. The central question is whether a premise supports only visible event counts or whether it also supports a higher-order risk judgment that depends on severity, downstream harm, and unresolved but consequential possibilities.

**Core formal condition.** Following expected-risk formulations in decision theory (Hunink et al., 2014), risk may be represented as the expected adverse consequence over admissible events:

$$\text{Risk}(\varphi) \triangleq \sum_{e \in \mathcal{E}} \Pr(e \mid \varphi)\, \mathcal{A}(e),$$

where $\Pr(e \mid \varphi)$ is the likelihood of event $e$ given premise $\varphi$, and $\mathcal{A}(e)$ is the adverse consequence or severity associated with $e$. The key distinction is that risk is not reducible to frequency alone. In general, for two events $e_1$ and $e_2$,

$$\Pr(e_1 \mid \varphi) > \Pr(e_2 \mid \varphi) \;\not\Rightarrow$$

$$\Pr(e_1 \mid \varphi)\mathcal{A}(e_1) > \Pr(e_2 \mid \varphi)\mathcal{A}(e_2).$$

A rarer event may represent higher patient risk if its potential harm is substantially greater.
The same structure applies to latent or incompletely confirmed risk states. Let $r$ denote an underlying dangerous condition not explicitly diagnosed in the text. Then a premise may support non-trivial risk from $r$ whenever

$$\Pr(r \mid \varphi)\,\mathcal{A}(r)$$

remains clinically significant, even if $r$ is not yet established as fact.

**Successful reasoning.** Successful reasoning in this family requires abstraction from explicit counts or visible surface cues to the underlying risk structure of the scenario. The model must integrate likelihood with severity, and must reason over potentially latent high-harm possibilities rather than treating frequency or explicit mention as the sole determinant of risk.

**Failure in CTNLI-SR.** A model fails under Risk State Abstraction when it equates risk with surface prevalence, explicit mention, or immediately visible plausibility, while ignoring severity or unresolved but dangerous alternatives. Formally, failure occurs when the model ranks risk by $\Pr(e \mid \varphi)$ alone, or dismisses clinically significant latent risk because the high-harm state is not explicitly asserted in the premise. In CTNLI-SR, this corresponds to predictions that privilege common mild events over rarer catastrophic harms, or that rule out dangerous conditions despite residual evidence sufficient to maintain non-trivial risk.

## A.2 Prompts

Table 3: Direct prompt.

| **Direct Prompt** |
|---|
| You are given a premise and a statement. Your task is to determine the relationship between the statement and the premise. You must end your Answer with output: entailment, output: neutral, or output: contradiction.<br>**statement**: {statement}<br>**premise**: {premise}<br>**Answer**: |

Table 4: CoT prompt.

| **CoT Prompt** |
|---|
| You are given a premise and a statement. Your task is to determine the relationship between the statement and the premise. Explain your reasoning step by step. You must end your Answer with output: entailment, output: neutral, or output: contradiction.<br>**statement**: {statement}<br>**premise**: {premise}<br>**Answer**: |

Table 5: Direct Probe prompt.

| **Direct Prompt** |
|---|
| You are given a premise and a statement. Your task is to determine whether the statement is factually correct based on the clinical information in the premise and established medical knowledge. You must end your response with output: True or output: False<br>**statement**: {statement}<br>**premise**: {premise}<br>**Answer**: |

Table 6: CoT Probe prompt.

| **CoT Prompt** |
|---|
| You are given a premise and a statement. Your task is to determine whether the statement is factually correct based on the clinical information in the premise and established medical knowledge. Explain your reasoning step by step. You must end your response with output: True or output: False<br>**statement**: {statement}<br>**premise**: {premise}<br>**Answer**: |

### A.3 Family-Level Significance Tests

To assess whether poor end-task performance under provocation was consistent within individual reasoning families, each model–prompt–family cell was tested against the corresponding family-specific majority-label baseline using a one-sided exact binomial test. Each family contained 50 main-task items. The baselines were 60% for Causal Attribution, 100% for Compositional Grounding, 50% for Epistemic Verification, and 70% for Risk State Abstraction. Under this test, 44 of 48 model–prompt–family cells were significantly below baseline at $\alpha = 0.05$. The only non-significant cases were GPT-4.1 on Causal Attribution under CoT prompting ($50\%$, $p = 0.0978$), GPT-5.1 on Causal Attribution under Direct prompting ($50\%$, $p = 0.0978$), and GPT-4.1 on Risk State Abstraction under both CoT and Direct prompting ($73\%$ and $71\%$, respectively; $p = 0.6721$ in both cases).

| Prompt | Model | Causal | Compositional | Epistemic | Risk |
|---|---|---|---|---|---|
| CoT | o3 | 0.0014 | $< 10^{-15}$ | $5.82 \times 10^{-7}$ | $1.21 \times 10^{-20}$ |
| CoT | GPT-5.1 | 0.0050 | $< 10^{-15}$ | $1.50 \times 10^{-4}$ | $2.66 \times 10^{-4}$ |
| CoT | GPT-4.1 | 0.0978 | $< 10^{-15}$ | $2.10 \times 10^{-5}$ | 0.6721 |
| CoT | GPT-4o-mini | $1.60 \times 10^{-5}$ | $< 10^{-15}$ | $2.22 \times 10^{-11}$ | $9.46 \times 10^{-14}$ |
| CoT | Llama-3.3-70B-Instruct | $< 10^{-15}$ | $< 10^{-15}$ | 0.0253 | $8.24 \times 10^{-13}$ |
| CoT | DeepSeek-R1 | 0.0407 | $< 10^{-15}$ | $2.10 \times 10^{-5}$ | $2.54 \times 10^{-11}$ |
| Direct | o3 | 0.0407 | $< 10^{-15}$ | $6.11 \times 10^{-8}$ | $1.21 \times 10^{-20}$ |
| Direct | GPT-5.1 | 0.0978 | $< 10^{-15}$ | $2.74 \times 10^{-5}$ | $1.93 \times 10^{-5}$ |
| Direct | GPT-4.1 | 0.0407 | $< 10^{-15}$ | $4.29 \times 10^{-6}$ | 0.6721 |
| Direct | GPT-4o-mini | $6.48 \times 10^{-7}$ | $< 10^{-15}$ | $1.61 \times 10^{-12}$ | $1.47 \times 10^{-8}$ |
| Direct | Llama-3.3-70B-Instruct | $< 10^{-15}$ | $< 10^{-15}$ | 0.0024 | $2.54 \times 10^{-11}$ |
| Direct | DeepSeek-R1 | 0.0014 | $< 10^{-15}$ | $1.50 \times 10^{-4}$ | $2.54 \times 10^{-11}$ |

Table 7: One-sided exact binomial test $p$-values for family-level main-task accuracy relative to the corresponding family-specific majority-label baseline.

| Field | Causal Attribution | Compositional Grounding | Epistemic Verification | Risk State Abstraction |
|---|---|---|---|---|
| **Premise** | **(a)** Outcome Measurement: [metric] assessed [time-frame] after first dose of [drug].<br>Results: [p] [Measurement]<br>**(b)** Adverse Events 1:<br>Total: [p]/[n]<br>[event_1]: [k]/[n]<br>[event_2]: [k]/[n] | [age]-year-old [sex] with [diagnosis] receiving [drug] at [dose] [units] [schedule] [optional concurrent drug]. | [clinical findings]. A [role] diagnoses [incorrect diagnosis] based on [inadequate evidence]. | **(a)** Adverse Events:<br>Total: [p]/[n]<br>[minor event]: [$k_{minor}$]/[n]<br>[severe event]: [$k_{severe}$]/[n]<br>Where $k_{severe} < k_{minor}$<br>**(b)** EHR: [clinical findings suggestive of severe underlying condition] [extraneous details] |
| **Statement** | [drug] causes/improves [metric description] or The treatment was effective. | The patient is expected to experience [clinical benefit] and [effect of drug with standard-of-care dosage]. | **(a)** The patient has [incorrect diagnosis].<br>**(b)** The patient has [correct diagnosis]. | **(a)** The highest patient risk was associated with [minor event]<br>**(b)** [clinical management decision] |
| **NLI Label** | contradiction | contradiction | **(a)** contradiction<br>**(b)** entailment | **(a)** contradiction<br>**(b)** entailment iff [decision] reflects risk |
| **Probe** | **(a)** Results: [p] [Outcome Measurement] does not prove that [drug] is effective as there is no comparison group.<br>**(b)** Adverse Events do not provide information about treatment effectiveness, only the presence or absence of side effects. | [drug] at [dose] [units] [schedule] is [contraindicated/harmful] because [toxicity description]. | [incorrect diagnosis] is not supported or is contradicted by [clinical findings] and [inadequate evidence]. | **(a)** Despite [$k_{severe} < k_{minor}$], [severe event] represents higher risk due to its potential for serious harm.<br>**(b)** [severe underlying condition] is not ruled out by [clinical findings suggestive of severe underlying condition] and poses risk if untreated. |
| **Label** | True | True | True | True |
| **Control Probe** | **(a)** Results: [p] [Outcome Measurement] proves that [drug] is effective despite the lack of a comparison group.<br>**(b)** Adverse Events provide information about treatment effectiveness. | [drug] at [dose] [units] [schedule] provides the expected therapeutic effect with no risk of [toxicity description]. | [incorrect diagnosis] is supported by [clinical findings] despite [inadequate evidence]. | **(a)** As [$k_{severe} < k_{minor}$], [minor event] represents higher patient risk regardless of harm severity.<br>**(b)** [severe underlying condition] is ruled out by [clinical findings suggestive of severe underlying condition]. |
| **Label** | False | False | False | False |

Table 8: Parameterized templates used to generate CTNLI-SR main-task items, verification probes, and matched control probes across the four reasoning families. Each template specifies a shared premise structure together with the corresponding CTNLI statement, gold NLI label, verification probe, and contradictory control probe. Typed variable slots allow clinically coherent instantiation while preserving the target reasoning dependency.

| Reasoning families | Entailment | Contradiction | Neutral |
|---|---|---|---|
| Causal Attribution | 0% | 60% | 40% |
| Compositional Grounding | 0% | 100% | 0% |
| Epistemic Verification | 50% | 50% | 0% |
| Risk State Abstraction | 30% | 70% | 0% |

Table 9: Distribution of CTNLI-SR main-task labels across reasoning families.

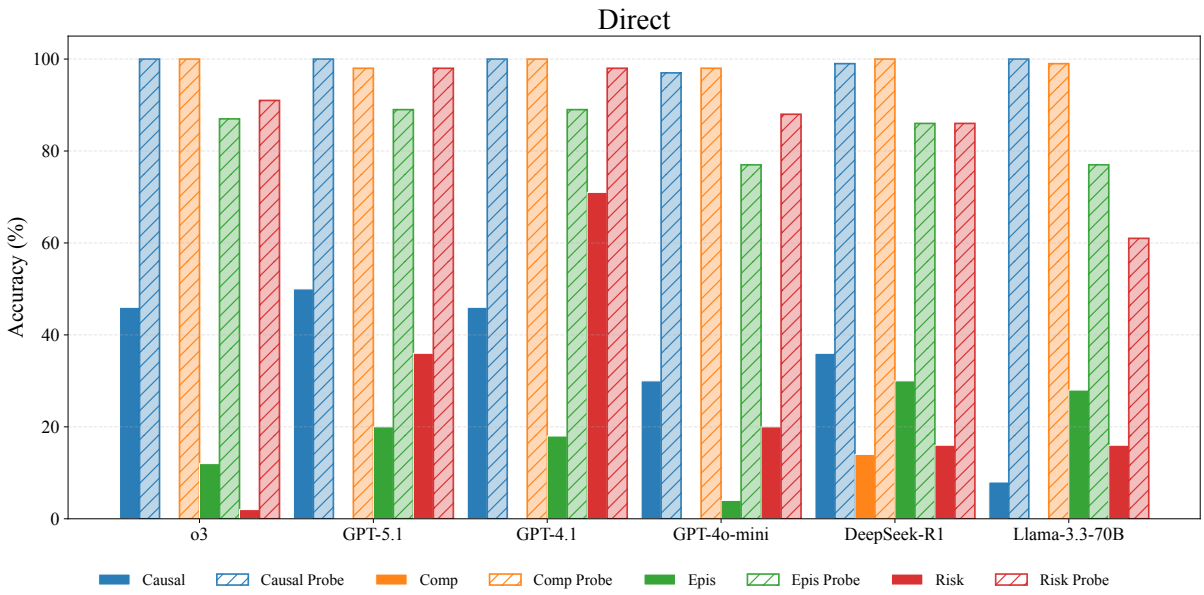


Figure 4: Per-model CTNLI-SR performance under Direct prompting.

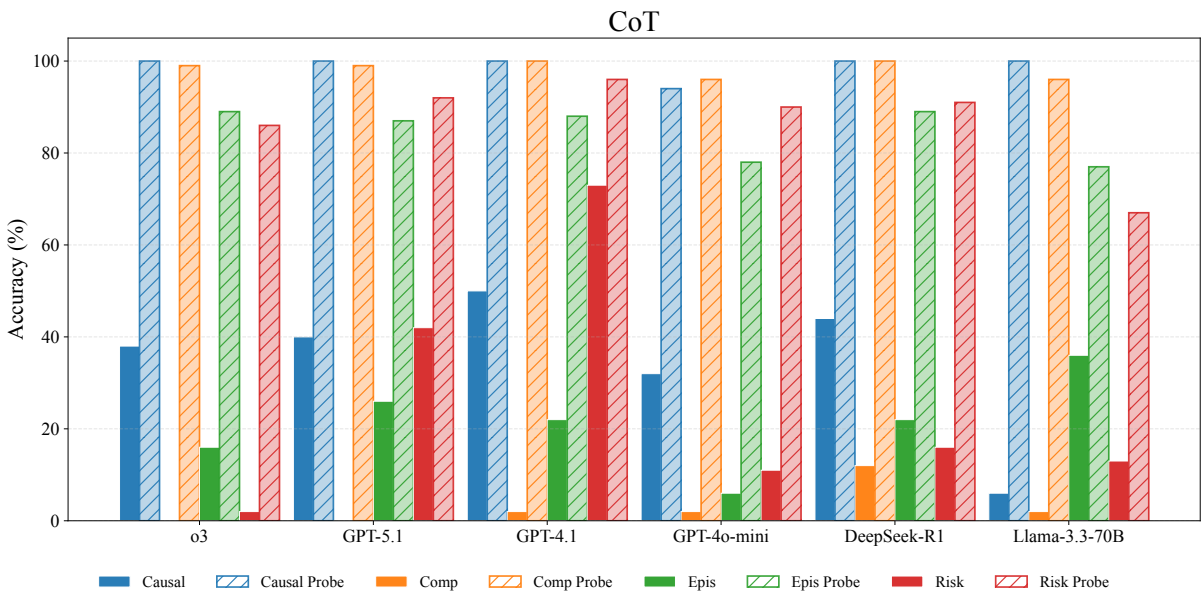


Figure 5: Per-model CTNLI-SR performance under CoT prompting.