Dissecting Clinical Reasoning in Natural Language Inference for Large Language Models

Maël Jullien[1,3], Marco Valentino[4], Leonardo Ranaldi[5], 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
[5] School of Informatics, University of Edinburgh, UK
[3] {firstname.surname}@idiap.ch

Abstract

Recent works on large language models (LLMs) have demonstrated the impact of prompting strategies and fine-tuning techniques on their reasoning capabilities. Yet, their effectiveness on *clinical natural language inference* (CTNLI) remains underexplored. This study presents the first controlled evaluation of how prompt structure and efficient fine-tuning jointly shape model performance in CTNLI.

We inspect four classes of prompting strategies to elicit reasoning in LLMs at different levels of abstraction, and evaluate their impact on a range of clinically motivated reasoning types. For each prompting strategy, we construct high-quality demonstrations using a frontier model to distil multi-step reasoning capabilities into smaller models ($\leq$4B parameters) via Low-Rank Adaptation (LoRA). Across different LLMs fine-tuned on the NLI4CT benchmark, we found that prompt type alone accounts for up to 44% of the variance in macro-F_1. Moreover, LoRA fine-tuning yields consistent gains of +8–12 F_1, raises output alignment above 97%, and narrows the performance gap to GPT-4o-mini to within 7.1%. Additional experiments on reasoning generalisation reveal that LoRA improves performance in 75% of the models on MedNLI and TREC Clinical Trials.

Overall, these findings demonstrate that (i) prompt structure is a primary driver of clinical NLI reasoning performance, (ii) compact models equipped with strong prompts and LoRA can rival frontier-scale systems, and (iii) reasoning-type-aware evaluation is essential to uncover prompt-induced trade-offs. Our results highlight the promise of combining prompt design and lightweight adaptation for more efficient and trustworthy clinical NLP systems, providing insights on the strengths and limitations of widely adopted prompting and parameter-efficient techniques in specialised domains[1].

[1]All code, annotations, prompts, demonstrations, and checkpoints will be released upon publication.

1 Introduction

LLMs now achieve state-of-the-art performance across a broad spectrum of reasoning tasks, including mathematics, commonsense, and science (Brown et al., 2020; Bommasani et al., 2021; Achiam et al., 2023). However, LLM performance is closely correlated to prompt design. A growing body of work has shown that different prompting strategies, such as chain-of-thought (Wei et al., 2022), self-ask (Press et al., 2022), and ReAct (Yao et al., 2023), can yield significant variation in reasoning behaviour (Zhang et al., 2022; Wen et al., 2025; Mondorf and Plank, 2024).

LLMs now achieve state-of-the-art performance across a broad spectrum of reasoning tasks, including mathematics, commonsense, and science (Brown et al., 2020; Bommasani et al., 2021; Achiam et al., 2023). However, LLM performance is closely correlated to prompt design. A growing body of work has shown that different prompting strategies, such as chain-of-thought (Wei et al., 2022), self-ask (Press et al., 2022), and ReAct (Yao et al., 2023), can yield significant variation in reasoning behaviour (Zhang et al., 2022; Wen et al., 2025; Mondorf and Plank, 2024).

Prompt engineering has been applied to clinical tasks such as medical question answering (Nori et al., 2023), scientific claim verification (Ma et al., 2023), and natural language inference (e.g., NLI4CT) (Jullien et al., 2023a). Prior work on NLI4CT and related shared-task settings has documented many of the clinically specific challenges of the benchmark, including difficulties arising from protocol language, eligibility criteria, and domain-specific reasoning (Jullien et al., 2023b, 2024, 2023a). However, these studies also vary substantially in model architecture, prompting strategy, and tuning protocol, which makes it difficult to isolate how inference structure itself affects performance. As a result, it remains unclear how prompt-

Findings of the Association for Computational Linguistics: ACL 2026, pages 26229–26250

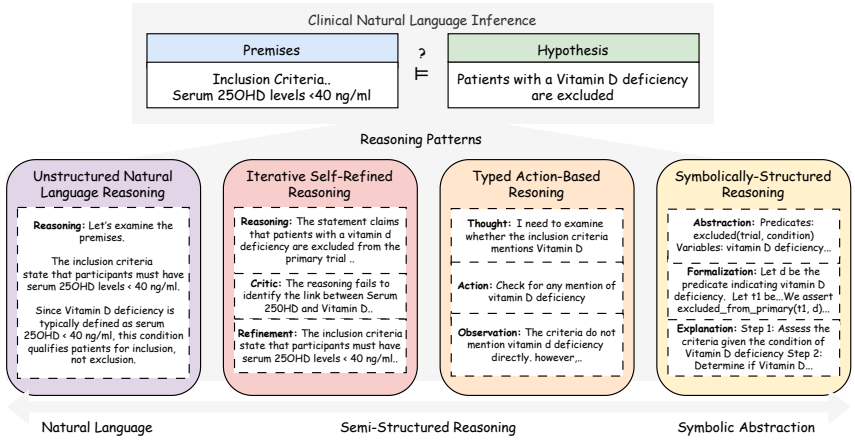

Figure 1: Example reasoning trajectories for a single NLI4CT instance under four prompting strategies: NLR, ISRR, TAR, and SSR. The prompts are shown in the context of our five-stage experimental framework: **(1)** Reasoning types defined and annotated; **(2)** Prompt categories defined and instantiated as structured scaffolds; **(3)** Generation of high-precision demonstrations; **(4)** Compact models adapted via LoRA on prompt-aligned data; **(5)** Evaluation performed by reasoning type and benchmark.

ing and parameter-efficient adaptation interact with the composite reasoning demands of the task under controlled conditions.

This paper presents the first systematic evaluation of prompting and parameter-efficient adaptation for CTNLI under controlled conditions. By holding model scaling ($\geq$4B), and training configuration, our experimental setup enables direct comparison across five structurally distinct prompting paradigms (Figure 1). To support interpretable analysis, we introduce expert-labeled reasoning annotations for the NLI4CT test set, spanning six categories: Clinical, Lexical Equivalence, World Knowledge, Expectation-Driven Evidence, Quantitative Comparison, and Quantitative Derivation.

Our experiments reveal the following key findings:

Prompt structure is a primary driver of clinical NLI reasoning performance. Prompt structure alone explains up to **44 %** of the variance in F_1. This reveals the importance of prompt design for CTNLI, and, at the same time, the persisting sensitivity of LLMs to specific input instructions.

Compact models equipped with strong prompts and LoRA can rival frontier-scale systems. LoRA supplies consistent gains of **+8–12 F_1** and lifts answer validity above 97%, allowing a 3.8B model to trail GPT-4o-mini by 7.1%. This shows that parameter-efficient adaptation techniques using high-quality demonstrations are a viable solution to reduce the gap between frontier proprietary models and smaller open-source models in specialised domains.

Reasoning-type-aware evaluation is essential to uncover prompt-induced trade-offs. Prompt strategy explains 30% to 44% of variance across reasoning types, when controlling for model architecture and LoRA. This demonstrates the importance of fine-grained reasoning evaluation when assessing inference strategy in the clinical domain.

LoRA fine-tuning can generalise to different CTNLI tasks. We found significant improvements on MedNLI and the TREC Clinical Trials Track when fine-tuning on NLI4CT alone, where LoRA boosts F_1 in 75 % of model–prompt settings. Moreover, we found that the gap between fine-tuned models and GPT-4 is particularly reduced on more complex NLI tasks (i.e., TREC). This demonstrates that some of the inference capabilities acquired via parameter-efficient fine-tuning are preserved across distribution and complexity variations of the CTNLI tasks.

2 Methodology

2.1 Methodology Overview

This study follows a five-stage framework illustrated in Figure 1. **(1) Reasoning Type Definition.** A typology of six reasoning types essential for clinical inference is defined. Each instance in the NLI4CT test set is manually annotated with one or more reasoning labels. **(2) Prompt Strategy Design.** A taxonomy of four abstract prompt categories is defined, each representing a distinct

structural mode of inference. One representative prompt is constructed per category. **(3) Demonstration Collection.** For each prompt strategy, validated demonstrations are collected by applying GPT-4o-mini to the NLI4CT training set and retaining only instances for which the model prediction matches the gold label. **(4) LoRA Adaptation.** Four target models in the 1.5–3.8B parameter range are fine-tuned across all demonstration sets independently, using LoRA. **(5) Evaluation and Analysis.** All adapted models are evaluated on the full reasoning-annotated NLI4CT test set, as well as two out-of-domain generalisation benchmarks: MedNLI and TREC.

2.2 Reasoning Types in Clinical Trial NLI

Six reasoning categories are defined for NLI in this setting: *Clinical*, *Lexical Equivalence*, *Evidence*, *World Knowledge*, *Quantitative Comparison*, and *Quantitative Derivation*. Each category isolates a distinct inferential mechanism, that can be applied simultaneously, or concurrently within a single inference.

These categories are not intended as a universal taxonomy of clinical reasoning, but as an operationalisation of recurrent subskills identified in the clinical reasoning literature (Eva, 2005; Jay et al., 2024; Norman, 2024). They are defined at the level of granularity needed for reliable CTNLI annotation, and were also chosen for their prevalence in NLI4CT, ensuring sufficient representation for statistically meaningful analysis.

Table 1 presents examples for each category, which are formally defined below.

1. Clinical Trial Reasoning Application of clinical knowledge within the context of a clinical trial protocol. This reasoning interprets clinical concepts (labs, events, terminology) through the lens of operational definitions, regulatory frameworks, and trial procedures. As a result, common clinical terms and expressions often take on specialized meanings that diverge from general interpretations.

2. Lexical Equivalence Identifying semantically equivalent phrases, based on surface-level lexical synonymy.

3. Expectation-Driven Evidence Reasoning The statement is decomposed into a set of evidence markers, a discrete set of expectations about structural form, functional role, or factual content (e.g., actions, entities, measurements, or reported outcomes) that should appear in the premise if the statement holds. The premise is then examined for content that aligns with the expected form and function of these markers, permitting partial evidence. Under a closed-world assumption, the absence of the expected evidence is treated as evidence of negation.

4. World-Knowledge Inference Applying general world knowledge, and common sense rules.

5. Quantitative Comparison Direct numeric entailment by single step comparison of values and thresholds explicitly defined in the premise and statement.

6. Domain-Grounded Quantitative Derivation Multi-step arithmetic on variables that may be directly specified, contextually inferred, or estimated using domain knowledge.

2.3 Prompt Categories as Structural Abstractions

We introduce a taxonomy of four abstract prompt categories, each corresponding to a distinct structural mode of reasoning. These categories span a continuum from fully natural language to semi-symbolic abstraction. This taxonomy frames prompts as inductive biases that shape the reasoning trajectory by imposing structural scaffolds on how information is processed, decomposed, and justified. Definitions are provided in Table 2 and illustrated in Figure 1.

These are not proposed as an exhaustive theory of prompting, but as a minimal structural taxonomy over the prompting families instantiated in this study: linear verbal reasoning, reflective revision, interleaved reasoning-and-verification, and quasi-symbolic abstraction. This grouping is motivated by prior work showing that chain-of-thought, self-refinement, ReAct-style prompting, and quasi-symbolic decomposition induce different control flows and intermediate reasoning representations (Wei et al., 2022; Madaan et al., 2023; Yao et al., 2023; Ranaldi et al., 2025).

2.4 Automatic Demonstration Generation

Demonstrations are automatically generated by a frontier model applied to a source corpus. Predictions are compared to gold-standard labels, and only correctly classified instances are retained, yielding high-precision, label-faithful, and task-aligned demonstration sets.

Reasoning Type	Statement	Premise	Reasoning
Clinical Reasoning	One emesis episode and ondansetron given.	Vomiting was controlled.	Controlled = *no* vomiting and *no* rescue meds. Rescue therapy disqualifies control. $\Rightarrow$ Contradiction
Lexical Equivalence	Participants must have failed platinum therapy.	Inclusion Criteria: Previous treatment with Cis/Gem that didn't control disease	("Cis/Gem" $\Rightarrow$ "platinum-based regimens" & "Failed therapy" $\Rightarrow$ "disease not controlled") $\Rightarrow$ Entailment.
Expectation-Driven Evidence Reasoning	The primary trial has 2 separate cohorts	Intervention section: once-daily oral dose of empagliflozin 10mg	Expected cohort/phase names or distinct treatment paths; none found (closed-world assumption) $\Rightarrow$ Contradiction.
World-Knowledge Inference	The trial excludes children.	Only participants 18 years or older may enroll.	Age $\geq$ 18 implies participants are adults, thus excluding children $\Rightarrow$ Entailment.
Quantitative Comparison	Her CrCl is below 30 mL/min.	Creatinine clearance calculated as 28 mL/min.	$28 < 30 \Rightarrow$ Entailment.
Domain-Grounded Quantitative Derivation	All C1 patients receive higher ALT doses than C2 patients.	Cohort 1: ALT-801 0.04 mg/kg Cohort 2: ALT-801 0.01 mg/kg	Min C1 weight 40kg: $0.04 \times 40 = 1.6$mg. Max C2 weight 150kg: $0.01 \times 150 = 1.5$mg. $1.6 > 1.5 \Rightarrow$ Entailment.

Table 1: Examples of inference steps for each reasoning type on NLI4CT examples.

Category	Structural Signature	Reasoning Trajectory	Example Prompts
Unstructured Natural Language Reasoning (NLR)	Free-form explanation followed by a final answer, with no explicit phases or typed components.	Single contiguous segment: $r_1, r_2, \ldots, r_k \rightarrow y$	*"Let's think step by step"*, *"Explain your reasoning"* *"Describe your thought process."*
Iterative Self-Refined Reasoning (ISRR)	Multi-stage reasoning with internal self-evaluation. An initial response, a critique or verification, and a revised answer.	Stage 1: y' (draft) Stage 2+: c (critique) Stage 3+: y (revised)	*"Verify your reasoning at every step"*, *"Self-debate between supportive and critical perspectives."*
Typed Action-Based Reasoning (TAR)	Natural language interleaved with discrete actions (e.g., `SEARCH`, `LOOKUP`) that return observable results.	$(\text{Thought}_t, \text{Action}_t, \text{Obs}_t)_{t=1}^{T} \rightarrow y$	*"Use Search_PUBMED(query) to look up..."*, *"Use QUERY_DB(sql) to retrieve ..."*
Symbolically Structured Reasoning (SSR)	A symbolic representation (e.g., predicate logic, code, or JSON object) before deriving an answer.	$X \xrightarrow{\text{Abstract}} S \xrightarrow{\text{Solve}} y$, where S belongs to a formal language	*"Translate this to predicate logic"*, *"Write a python function to solve.."*

Table 2: Illustration of reasoning trajectories across four prompting categories (NLR, TAR, ISRR, SSR), including their structural signatures, control flow patterns, and representative prompt examples.

2.5 LoRA Fine-Tuning

Parameter-efficient adaptation is performed using LoRA (Hu et al., 2022), which injects trainable low-rank adapters into the attention and feed-forward layers while keeping all other parameters fixed. This enables consistent, efficient fine-tuning across models and prompt strategies, preserving core model representations and facilitating controlled evaluation of prompt sensitivity and generalisation. Fine-tuning is supervised using a fixed number of randomly sampled demonstrations per prompt type, selected to balance convergence stability and overfitting risk. A uniform hyperparameter configuration is applied across models for comparability.

3 Empirical Evaluation

3.1 Representative Prompting Strategies

We instantiate one representative prompt per abstract category defined in Table 2: Chain-of-Thought (NLR), Self-Critique (ISRR), ReACT (TAR), and QuaSAR (SSR), along with a zero-shot baseline. Each prompt reflects the structural pattern and reasoning mode characteristic of its respective category. Full prompt definitions, templates, and examples are provided in Section A.3 , with representative reasoning trajectories illustrated in Figure 1.

3.2 Dataset Selection

NLI4CT The NLI4CT dataset (Jullien et al., 2023a), defines an NLI task over structured sections of clinical trial reports (CTRs) from `ClinicalTrials.gov`. Given a natural language *statement*, and a corresponding *CTR premise*, a passage drawn from one of four sections: *Eligibility*, *Intervention*, *Results*, or *Adverse Events*, predict whether the premise entails or contradicts the statement by assigning one of two labels: *Entailment* or *Contradiction*. Each NLI4CT test instance was expert-annotated with one or more of six reasoning categories, indicating the types of reasoning needed to solve it (Full details in Section A.4).

Generalisation Datasets To evaluate cross-domain generalisation, we test on two external CTNLI datasets: MedNLI (Romanov and Shivade, 2018) and the TREC 2022 Clinical Trials Track (Roberts et al., 2022). MedNLI provides sentence-level entailment labels derived from MIMIC-III patient notes, while TREC frames eligibility inference over synthetic patient-trial pairs. We use balanced 400-instance subsets from each dataset, reformulated as binary NLI tasks (See Section A.5 for dataset details).

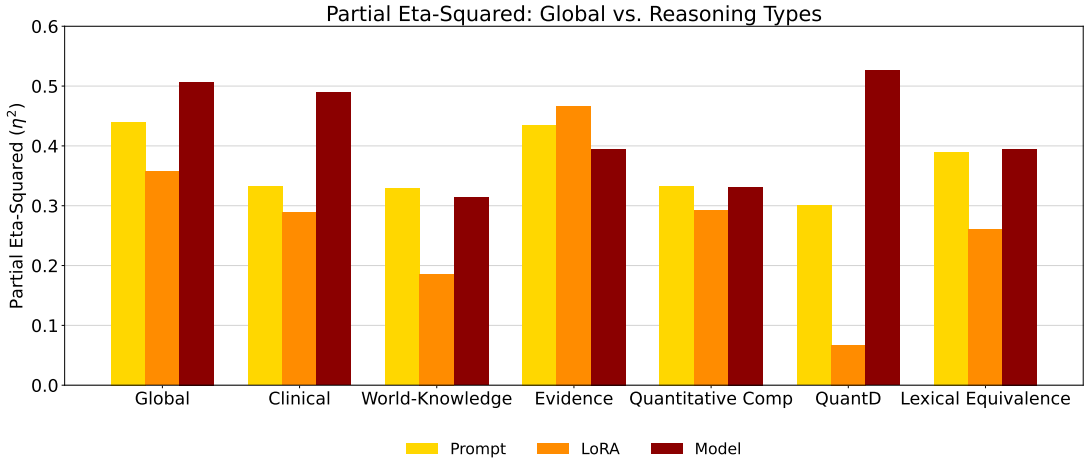

Figure 2: Partial η^2 values showing the proportion of variance in F_1 performance explained by Prompt, LoRA adaptation, and Model architecture, both globally and across reasoning types.

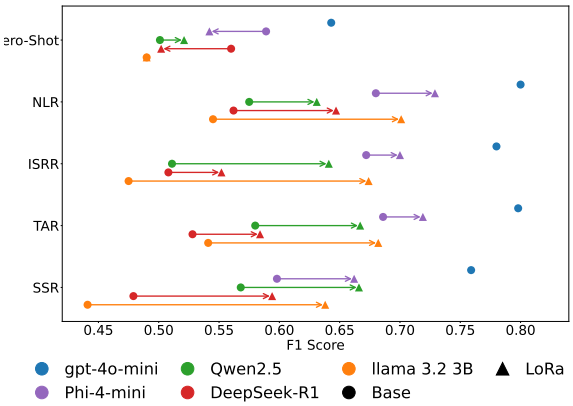

Figure 3: Macro F_1 Scores on the NLI4CT Test Set

3.3 Demonstration Generation

Demonstrations are generated with GPT-4o-Mini (OpenAI, 2024) on 1,900 instances from the combined NLI4CT training and development sets for each prompt type. Only correctly classified instances are retained to construct high-precision demonstration sets (Table 24). We additionally assess the reliability of the reasoning-type annotations used in our analysis via an inter-annotator agreement study (Fleiss' $\kappa = 0.68$ for primary-type assignment; Appendix A.6).

3.4 Model Selection

We evaluate four instruction-tuned LLMss: `LLaMA-3.2-3B` (Grattafiori et al., 2024), `Qwen-2.5-3B-Instruct` (Yang et al., 2024), `DeepSeek-R1-Distill-Qwen-1.5B` (DeepSeek-AI, 2025), and `Phi-4-Mini-Reasoning-3.8B` (Xu et al., 2025). This selection of smaller models reflects the practical constraints of real-world clinical settings. Regulatory and infrastructure limitations often necessitate on-premise deployment, where sub-4B models offer a tractable trade-off between computational cost, memory footprint, and task performance (See Section A.1 for details).

3.5 Fine-Tuning Details

Each model is fine-tuned with LoRA using 500 randomly selected demonstrations corresponding to each prompt type. A uniform hyperparameter configuration is applied across all models to ensure consistency, with the exception of Phi-4, which requires a distinct target module configuration due to architectural constraints. (Full details are provided in Section A.2).

3.6 Fine-Tuning Details

Each model is fine-tuned with LoRA using 500 randomly selected demonstrations corresponding to each prompt type. A uniform hyperparameter configuration is applied across all models to enable controlled comparison, with the exception of Phi-4, which requires a a minor configuration change due to architectural constraints. This design prioritises comparability and variance attribution over condition-specific optimisation. Full training details are provided in Section A.2.

4 Results on NLI4CT

Results Overview Overall, the results on NLI4CT reveal three central factors shaping performance on CTNLI: prompt strategy, reasoning type sensitivity, and LoRA-based fine-tuning. Controlling for model identity and LoRA status, prompt strategy explains **44 %** of the variance in macro–F_1

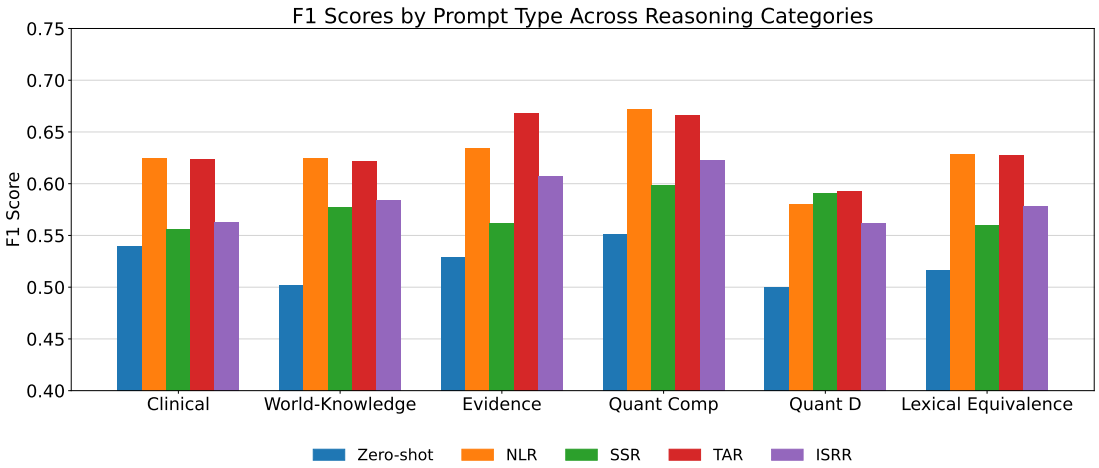

Figure 4: Performance by Reasoning Type on the NLI4CT Test Set (Macro F_1)

on NLI4CT and **30–44 %** within each individual reasoning class. Regarding the prompting strategy, we found that NLR and TAR yield the highest average F_1 scores. However, different prompts induce distinctive precision–recall profiles: ISRR maximises recall, whereas SSR maximises precision. Additionally, the results indicate that prompt strategies redistribute model strengths and weaknesses across reasoning types, rather than yielding uniform performance gains. Regarding the inference types, we found that *Evidence* and *Quantitative Comparison* are the least challenging (mean $F_1 \geq 0.653$), whereas *Quantitative Derivation* remains the most difficult, with the lowest average performance overall ($F_1 = 0.565$, 95% bootstrap CI [0.51, 0.62]) and only limited benefit from LoRA, while also showing strong dependence on model size. Finally, we observe that LoRA fine-tuning consistently improves macro–F_1 by **+8–12** % for every prompt and model type except zero-shot, increasing also answer validity and format alignment above **97 %**.

4.1 Prompting Strategy and Reasoning Types

In this section, we analyse in more detail the performance and impact of the prompting strategies across different reasoning types on NLI4CT. The results are shown in Figure 2, 3 and 4, Tables 25, 22, 27, 16 and 26.

Prompt strategy explains 44% of the variance in F_1 performance, after controlling for the effects of model identity and LoRA. Using a fixed-effects Type II ANOVA controlled for model architecture and LoRA adaptation (shown in Table 25, Figure 2). Prompt strategy captures 44 % of the variance left unexplained by the other two factors ($\eta^2_{\text{partial}} = 0.440$, $p < 0.001$). This effect is statistically significant and comparable to model architecture (51%) and greater than LoRA (36%). This confirms that prompt structure is a primary lever on performance.

NLR and TAR achieve the highest average F_1 across all models, regardless of size or fine-tuning Across all model groups, and LoRA configurations, prompts rank consistently by average F_1 score. Specifically, NLR, TAR, ISRR, SSR, and Zero-shot, with the only deviation occurring in the base small model setting, where Zero-shot slightly outperforms SSR (Figure 3). This consistency suggests that the effectiveness of prompt methods is invariant to both model scale and LoRA. SSR and ISSR yield the lowest macro-F_1 scores in the base setting (Figure 3), though LoRA narrows the gap. This reflects the fact that SSR and ISSR impose greater planning overhead, longer-range dependency tracking, and tighter output constraints than NLR or TAR.

Prompt strategies enable precision–recall trade-offs As seen in Table 16, ISRR on GPT-4o-mini maximises recall (0.890) but sacrifices precision (0.648), while SSR inverses the trend (precision 0.832, recall 0.727). This suggests that prompt choice enables a tunable precision–recall trade-off.

Prompt strategy explains 30% to 44% of variance in F_1 across reasoning types, when controlling for model architecture and LoRA Fixed-effects Type II ANOVA per reasoning type shows prompt strategy explains 30–44% of F_1 variance ($\eta^2_{\text{partial}} = 0.3$–$0.435$, all $p < 0.022$), indepen-

dent of model and LoRA (Table 22, Figure 2). For World Knowledge and Quantitative Comparison, prompt design explained the largest share of variance. The average variance explained by prompt strategy within individual reasoning types is 35% compared to 44% overall, indicating that prompts reshape the distribution of strengths and weaknesses across reasoning types, rather than providing consistent gains overall.

Zero-shot performance confirms task difficulty and impact of adaptation Zero-shot F_1 scores for small models peak at 0.589 and average just 0.535, while GPT-4o-mini achieves only 0.643 (Figure 3). These results confirm that NLI4CT is a non-trivial task. Prompts that incorporate reasoning scaffolds, such as NLR and TAR consistently improve performance, with gains of up to +14 F_1 over the strongest zero-shot baselines.

NLR has the highest overall F_1, but TAR Leads on Evidence and Quantitative Derivation NLR achieves the highest mean F_1 across all reasoning classes (**0.660**, Table 27, Figure 4), only outperformed by TAR in the *Evidence*, and *Quantitative Derivation* classes. While NLR outperforms most alternatives, no single strategy is currently optimal across all reasoning types.

Evidence and quantitative comparison represent the least challenging reasoning types ($F_1 \geq$ 0.653) Across all models and prompting methods, *Quantitative Comparison* (**F_1 = 0.663**) and *Evidence* (**F_1 = 0.653**) emerge as the most tractable categories (Table 26, (Figure 4)). These tasks likely benefit from simpler entailment structures or more direct linguistic cues. In contrast, *Quantitative Derivation* remains the hardest, with a lower average F_1 and smaller variance in response to prompt or model changes.

SSR preferentially enhances quantitative reasoning SSR delivers the largest relative improvements on the two quantitative classes (Table 27). With LoRA adaptation, the mean macro-F_1 for *Quantitative Comparison* rises from 0.530 to 0.668 (+0.138, +26 %), and for *Quantitative Derivation* from 0.553 to 0.628 (+0.075, +14.7 %). No other prompt attains comparable fractional lift on these categories. While designed for formal semantic clarity, SSR's predicate logic structure appears well aligned with quantitative reasoning, likely due to its explicit decomposition of claims into tractable, logic-based subcomponents.

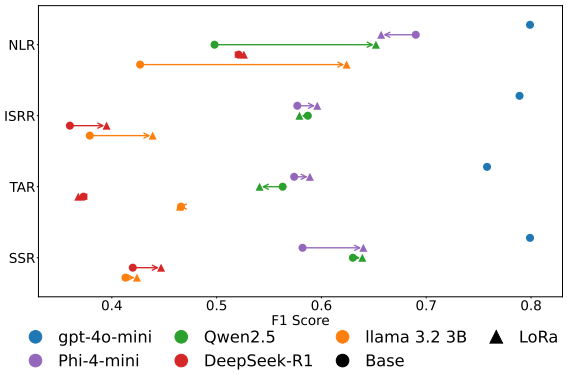

Figure 5: Macro F_1 Scores on the MedNLI Test Set

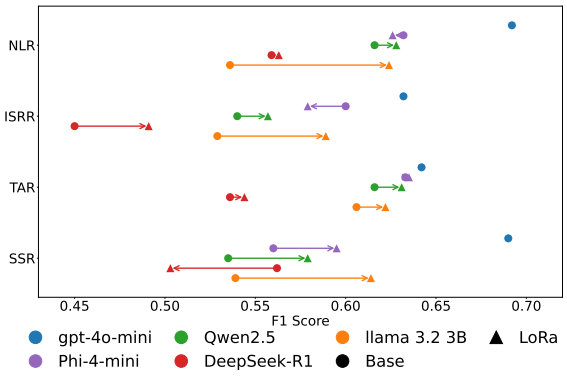

Figure 6: Macro F_1 Scores on the TREC Test Set

Practical Implications These findings establish prompt design as a key determinant of performance in CTNLI, with effects comparable to or exceeding those of model architecture and parameter-efficient adaptation. While NLR and TAR consistently yield the highest overall performance, no single prompt strategy optimally addresses all reasoning types. Prompt choice enables systematic trade-offs between precision and recall and reshapes the distribution of reasoning capabilities, with certain strategies conferring selective advantages on specific reasoning categories. Moreover, substantial variation in task difficulty across reasoning types suggests that prompt design should be aligned with the inferential demands of the target application. Future work in CTNLI should prioritise the identification of dominant reasoning types and the development of prompts tailored to their specific requirements.

4.2 The Impact of LoRA Fine-Tuning

LoRA fine-tuning improves F_1 by +8–12%for All Prompts Except Zero-Shot On average LoRA yields substantial gains in F_1 performance of +8-12% for all prompts, except Zero-Shot that loses 2 % (Figure 3) due to the absence of intermediate

supervision signals. The highest relative F_1 gain from LoRA observed with SSR on LLaMA 3.2 3B with a +44.7% relative improvement (Figure 3). This disparity between Zero-Shot and SSR highlights a synergy between fine-tuning and prompt structure: parameter adaptation is most beneficial when paired with prompts that elicit interpretable reasoning trajectories. Notably, the relative ranking of prompt types is already visible before fine-tuning and remains largely preserved after adaptation, indicating that LoRA amplifies or stabilises existing differences rather than creating them.

LoRA fine-Tuning raises answer validity to over 97% across all models Validity is defined as a single, correctly formatted label from the task-defined set (i.e., *entailment* or *contradiction*). LoRA-tuned models generate well-formed outputs, exceeding 97% validity across all configurations (Table 16). In contrast, base models are prone to formatting errors or incomplete responses, for instance, DeepSeek-base with ISRR produces valid outputs only 70.6% of the time.

LoRA most benefits structured semantic reasoning, but struggles with layered arithmetic reasoning LoRA improves F_1 on all reasoning types, with the largest gains in structured semantic reasoning (Figure 8). Averaged across all prompt methods, largest F_1 gains are *Evidence* (+0.117), *Quantitative Comparison* (+0.096), and *Lexical Equivalence* (+0.092) (Table 27). By contrast, *Quantitative Derivation*, which saw the lowest lift (+0.047) requires additional layers of symbolic reasoning not easily optimized through text-based supervision (Nye et al., 2021; Wei et al., 2022). These results imply that LoRA is more effective at reinforcing shallow symbolic patterns than deeper computational abstractions (Hu et al., 2022).

Model size still affects F_1, but best small model trails GPT-4o-mini by only 7.1% GPT-4o-mini maintains a consistent performance lead across all prompts, achieving a peak F_1 score of 0.8 with NLR (Figure 3). However, Phi-4-LoRA with NLR, reaches an F_1 of 0.729, within 7.1% of GPT-4o-mini, despite a significantly lower parameter count and computational requirements.

Practical Implications LoRA offers an efficient and scalable approach to improving CTNLI performance. Consistent F_1 gains across reasoning types and validity rates exceeding 97% indicate improvements in both predictive accuracy and output reliability. Moreover, prompt-based LoRA tuning allows compact models such as Phi-4 to approach the performance of frontier systems like GPT-4o-mini, providing a cost-effective alternative. However, Quantitative Derivation remains the hardest reasoning type and an important open challenge, with the smallest gains under adaptation. This suggests that text-only supervision has limited capacity to resolve numerically grounded clinical reasoning failures, and that practically reliable performance may require hybrid methods integrating LLMs with symbolic or other structured computational support. A natural direction for future work is to test whether these variance trends persist at larger model scales.

5 Generalisation to Different NLI Tasks

LoRA improves F_1 in 75% of cases (24/32), with gains up to +20% Across the MedNLI and TREC datasets, LoRA-tuning on NLI4CT improves F_1 in 24/32 model–prompt combinations (Figures 5, 6, Table 15, 14). Gains range from increases of +0.01 to +0.20, with no degradation exceeding –0.06, indicating that LoRA tuning is generally safe and beneficial for improving generalisation across model–prompt pairs.

LoRA preserves structural validity on different tasks LoRA is capable of enforcing output structure and reliability on tasks where the models are not explicitly trained on. All LoRA-tuned models produce 100% valid outputs on MedNLI and $\geq$94.5% on TREC. While F_1 increases are mostly recall-driven (e.g., Qwen2.5-NLR recall: 0.667 to 0.725), precision generally stays flat or declines slightly. This suggests that LoRA contributes to recovering more true positives by regularizing output format (Figures 5 and 6).

LoRA reduces the gap between models on complex tasks. Interestingly, we found that LoRA contribute to substantially reducing the gap between the fine-tuned model and frontier models on TREC, the task that is generally the most challenging for all the tested models. As shown in Figure 6, even if not explicitly trained on TREC, smaller models equipped with LoRA fine-tuning on NLI4CT can achieve performances that are comparable with GPT-4o-mini.

Practical Implications LoRA tuning on NLI4CT produces models that generalise reliably across prompt-model variants, and CTNLI tasks with consistent gains in recall and output validity.

This indicates the potential of LoRA strategies to effectively deploy small open-source LLMs on specialised domains for practical applications.

6 Related Work

Prompt engineering has substantially improved LLM reasoning, through techniques such as Chain-of-Thought (CoT) (Wei et al., 2022), self-consistency decoding (Wang et al., 2022), TAR's iterative verification loops (Yao et al., 2023), and self-critique refinement (Madaan et al., 2023) have each improved performance on general reasoning benchmarks. In clinical NLP, Sci-CoT distils GPT–4 reasoning traces into compact models for scientific question answering (Ma et al., 2023), while MedPrompt steers GPT–4 with few-shot CoT and self-generated explanations, surpassing domain-specific systems on nine medical challenge sets (Nori et al., 2023). Despite these advances, systematic head-to-head comparisons of multiple prompting paradigms remain rare in general, and non-existent in the field of CTNLI.

Across two SemEval shared tasks (Jullien et al., 2023b, 2024), participants explored various prompting techniques for NLI4CT: zero-shot templates, few-shot in-context learning, CoT, contrastive CoT, and retrieval-augmented prompting. However, these strategies were trialled on heterogeneous base models, obscuring the extent to which observed gains originate from prompt design. Additionally, meta-analysis indicates that CoT offers the greatest benefit for mathematical and symbolic tasks (Sprague et al., 2024), but contributes less to the commonsense, knowledge-based, and pragmatic reasoning demanded by CTNLI.

Recent cross-domain studies have begun to isolate the effect of prompting on distinct reasoning categories (Table 28). ThinkPatterns-21k (Wen et al., 2025), Auto-CoT (Zhang et al., 2022), and Self-Ask (Press et al., 2022), for example, contrast several prompting patterns on mathematics, commonsense, and symbolic problems, while Beyond Accuracy (Mondorf and Plank, 2024) and Systematic Relational Reasoning (SRR) (Khalid et al., 2025) probe logical, causal, and spatial reasoning. Yet none of these efforts considers CTNLI.

Prior work has shown the potential of prompt-based reasoning in clinical NLP and surveyed prompt strategies across domains. What remains lacking is a controlled evaluation of how these strategies transfer specifically to CTNLI.

7 Conclusion

We present the first controlled study of prompting and parameter-efficient adaptation for CTNLI. We formalise four prompting paradigms and six clinically grounded reasoning categories, providing a framework for systematic analysis and dataset annotation. Our results show that prompt structure is a primary driver of performance: when controlling for model architecture and adaptation, it accounts for up to 44 % of the macro-F_1 variance on NLI4CT. Parameter-efficient fine-tuning further enhances compact models, with LoRA improving 3–4 B models by 8–12 % in macro-F_1, increasing answer validity above 97 %, and reducing the gap to GPT-4o-mini to 7.1 %. Reasoning-aware evaluation reveals trade-offs, with prompt choice explaining 30–44 % of the variation across reasoning categories. Finally, LoRA demonstrates strong generalisation, improving F_1 in 75 % of model–prompt pairs on MedNLI and TREC-CT.

8 Limitations

Despite the encouraging gains reported, several factors constrain the scope and generalizability of our findings.

Synthetic Demonstrations. All demonstrations were generated with GPT-4o-Mini. This risks importing factual or stylistic biases present in the teacher model into the fine-tuned students. Future work should explore human-curated or adversarially filtered demonstration pools to mitigate inherited bias.

Clinical Readiness. The models are trained and evaluated solely on benchmark datasets and have not undergone validation on real patient data, error-sensitivity analyses, or regulatory review. Consequently, these systems are *not* suitable for clinical use. Additionally, deploying NLI models in healthcare raises additional ethical duties, including safeguarding patient privacy, preventing harmful automation bias, and ensuring transparency of decision support. These issues are beyond the scope of this study.

Model Scale. We restrict our experiments to checkpoints below 4 B parameters. Larger models may interact differently with the same prompt strategies, potentially altering the utility of LoRA, or certain prompt strategies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kevin W. Eva. 2005. What every teacher needs to know about clinical reasoning. *Medical Education*, 39(1):98–106.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Robert Jay, Clare Davenport, and Rakesh Patel. 2024. Clinical reasoning—the essentials for teaching medical students, trainees and non-medical healthcare professionals. *British Journal of Hospital Medicine*, 85(7):1–8.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Maël Jullien, Marco Valentino, and André Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1947–1962.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023a. 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.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023b. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226.

Irtaza Khalid, Amir Masoud Nourollah, and Steven Schockaert. 2025. Benchmarking systematic relational reasoning with large language and reasoning models. *arXiv preprint arXiv:2503.23487*.

Yuhan Ma, Chenyou Fan, and Haiqi Jiang. 2023. Scicot: Leveraging large language models for enhanced knowledge distillation in small models for scientific qa. In *2023 9th International Conference on Computer and Communications (ICCC)*, pages 2394–2398. IEEE.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Geoffrey Norman. 2024. Dual process models of clinical reasoning: The central role of working memory capacity. *Medical Education*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, and 1 others. 2021. Show your work: Scratchpads for intermediate computation with language models.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. Model card. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Leonardo Ranaldi, Marco Valentino, Alexander Polonsky, and Andrè Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. *arXiv preprint arXiv:2502.12616*.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and William R Hersh. 2022. Overview of the trec 2022 clinical trials track. In *TREC*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

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

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.

Pengcheng Wen, Jiaming Ji, Chi-Min Chan, Juntao Dai, Donghai Hong, Yaodong Yang, Sirui Han, and Yike Guo. 2025. Thinkpatterns-21k: A systematic study on the impact of thinking patterns in llms. *arXiv preprint arXiv:2503.12918*.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, and 1 others. 2025. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. 2025. Benchmarking reasoning robustness in large language models. *arXiv preprint arXiv:2503.04550*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

A Appendix

Uniform Training Protocol. A fixed set of hyper-parameters is applied to all model–prompt combinations for consistency. While this controls for tuning variance, it may inadvertently favour architectures/prompt strategies whose inductive biases align better with the chosen configuration.

A.1 Model Links

The following publicly available models were used in this study:

- **LLaMA-3.2–3B**: `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`
- **Qwen-2.5–3B-Instruct**: `https://huggingface.co/Qwen/Qwen2.5-3B-Instruct`
- **DeepSeek–R1–Distill–Qwen–1.5B**: `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`
- **Phi-4–Mini–Reasoning–3.8B**: `https://huggingface.co/microsoft/Phi-4-mini-reasoning`

A.2 Training Configuration

We fine-tuned our model using the Hugging Face `Trainer` API in conjunction with parameter-efficient fine-tuning (PEFT) via LoRA (Hu et al., 2022). Below, we detail the hyperparameters and training configuration used in our experiments.

A.2.1 LoRA Configuration

We applied LoRA to the attention and feed-forward layers of the model with the following settings:

- **Task type:** Causal Language Modeling (CAUSAL LM)
- **LoRA rank (r):** 8
- **LoRA scaling factor (α):** 16
- **LoRA dropout:** 0.1
- **Target modules:** `[q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj]`

- **Target modules for Phi:** `['down_proj', 'gate_up_proj', 'qkv_proj', 'o_proj']`
- **Bias:** None

A.2.2 Training Hyperparameters

The training was performed using the following hyperparameter settings:

- **Maximum sequence length:** 4096
- **Train batch size (per device):** 1
- **Eval batch size (per device):** 1
- **Gradient accumulation steps:** 16
- **Learning rate:** 2×10^{-5}
- **Weight decay:** 0.01
- **Warm-up steps:** 10
- **Maximum steps:** 500
- **Early stopping patience:** 2 evaluations

A.2.3 Optimization

The model selection criterion was the lowest evaluation loss on the validation set.

A.3 Prompt Descriptions & Examples

NLR: Chain–of–Thought (CoT). CoT instructs the model to verbalise step-by-step reasoning, breaking down complex problems into more manageable parts, and has been shown to improve performance over standard prompting methods, particularly in domains that require multi-step reasoning, such as mathematical problem-solving, commonsense inference, and logical deduction (Wei et al., 2022). See Table 5 and Figure 1 for an example of CoT reasoning on NLI4CT.

ISRR: Self-Critic. The Self-Critic prompt (Madaan et al., 2023) embeds an explicit reflection phase: the model drafts an answer, critiques its own reasoning, and then revises accordingly. This three-step workflow, exemplified in Table 6 and Figure 1, fosters error detection, reduces reasoning shortcuts, and has been linked to improved factual accuracy and safety in sensitive domains.

TAR: REACT (Reason & Act). REACT (Yao et al., 2023) alternates internal reasoning tokens (THOUGHT) with explicit operations on the environment (ACTION) followed by feedback (OBSERVATION), as shown in Table 3 and Figure 1. In this implementation, ACTION acts as a self-check that compares provisional inferences with the premise text. This iterative verify-before-commit loop has been shown to curb hallucinations in knowledge-intensive tasks and to encourage the retrieval of supporting evidence.

SSR: Quasi–Symbolic Abstract Reasoning (QuaSAR). QuaSAR (Ranaldi et al., 2025) decomposes a problem into four quasi-symbolic stages designed to tighten the model's reasoning trajectory: (i) Abstraction: identify predicates, variables, and constants; (ii) Formalisation: Using the abstracted variables, formalise the logical structure and constraints of the problem(s); (iii) Explanation: derive the solution over the formal representation; and (iv) Answering: translate the result back into natural language. These steps are demonstrated in Table 4 and Figure 1. This prompt enables explanatory arguments to be abstracted from specific world knowledge—for example, rather than learning 'The patient takes metformin, so they have diabetes,' the model learns the general pattern *"Taking [Drug X] implies [Condition Y]"* supporting generalisation across cases (Ranaldi et al., 2025).

Zero-Shot In the zero-shot setting, the model is prompted to produce a label directly, without any intermediate reasoning. This serves as a baseline or sanity check against which more structured strategies can be compared.

TAR Instructions
Statement: Patients with a Vitamin D deficiency are excluded
Premise: Inclusion Criteria..[46 words].. Serum 25OHD levels <40 ng/ml
Though: I need to examine whether the inclusion criteria mentions Vitamin D
Action: Check for any mention of vitamin D deficiency
Observation: The criteria do not mention vitamin d deficiency directly. however, ..[13 words]..
Though: The claim in the statement is supported by the premise. **Output: Contradiction**

Table 3: Examples of TAR reasoning

A.4 NLI4CT Description

NLI4CT is designed to support real-world clinical tasks and enable comprehensive evalua-

SSR Instructions
Statement: Patients with a Vitamin D deficiency are excluded
Premise: Inclusion Criteria..[46 words].. Serum 25OHD levels <40 ng/ml
Abstraction (s1): Predicates: excluded(trial, condition) Variables: vitamin D deficiency ..[23 words]
Formalisation (s2): Let d be the predicate indicating vitamin D deficiency. Let t_1 be the primary trial. We assert e_1: excluded_from_primary(t_1, d)..[48 words]
Explanation (s3): Step 1: Assess the criteria given the condition of Vitamin D deficiency Step 2: Determine if Vitamin D..[28 words]..
Answering (s4): Output: Contradiction

Table 4: Example of SSR reasoning

Chain of Though Instructions
Statement: Patients with a Vitamin D deficiency are excluded
Premise: Inclusion Criteria..[46 words].. Serum 25OHD levels <40 ng/ml
Reasoning: Let's examine the premises. The inclusion criteria state that participants must have serum 25OHD levels$<$40 ng/ml. Since Vitamin D deficiency is typically defined as serum 25OHD$<$40 ng/ml, this condition qualifies patients for inclusion, not exclusion. Therefore, the statement contradicts the premise. **Output: Contradiction**

Table 5: Example of CoT reasoning

ISRR Instructions
Statement: Patients with a Vitamin D deficiency are excluded
Premise: Inclusion Criteria..[46 words].. Serum 25OHD levels <40 ng/ml
Draft: The statement claims that patients with a vitamin d deficiency are excluded from the primary trial ..[41 words].. **Output: Entailment**
Critical Comments: The draft fails to identify the link between Serum 250HD and Vitamin D ..[52 words].. Additionally, the terminology for excluded should be consistent .
Final Response: The inclusion criteria state that participants must have serum 25OHD levels$<$40 ng/ml ..[37 words].. **Output: Contradiction**

Table 6: Examples of ISRR reasoning

Baseline Prompt
Determine whether the statement is an entailment or a contradiction of the premise. Answer only with either "entailment" or "contradiction". **Statement:** {statement} **Premise:** {premise} **Output:**

Table 7: Baseline prompt for NLI4CT

Chain of Thought Prompt
You are given a premise and a statement. Your task is to determine the relationship between the statement and the premise by analyzing them carefully. Instructions: • Carefully read the statement and the premise. • Think through your reasoning step by step, considering whether the statement logically follows from the premise or contradicts it. • Use chain-of-thought reasoning to reach your conclusion. • Based on your analysis, determine whether the statement is an 'entailment' or a 'contradiction' of the premise. • In your response, first provide your chain-of-thought reasoning, and then output 'entailment' or 'contradiction' as your final answer. • You must end your response with "output: entailment" or "output: contradiction". **Statement:** {statement} **Premise:** {premise} **Reasoning:**

Table 8: Chain-of-thought prompt for NLI4CT

QuaSAR Prompt
#Role You are an experienced expert skilled in answering complex problems through logical reasoning and structured analysis. #Task You are presented with an entailment problem that requires logical reasoning and systematic problem-solving. Given a statement and a premise, you are required to determine whether the statement follows from the premise. If the statement follows from the premise, end your response with "output: entailment". If the statement contradicts the premise, end your response with "output: contradiction". Please determine the entailment following these steps rigorously. #Steps 1) Please consider the following statement and premise and exemplify the relevant predicates, variables, and constants. Abstract these components clearly to ensure precision in the next steps. Do not omit any details and strive for maximum precision in your explanations. Refer to this step as Abstraction (s1) 2) For each predicate, variable and constant defined in s1, translate the statement and premise in formal symbolic representation. Please ensure that the formalisation captures the logical structure and constraints of the statement and premise. For clarity, provide the exact formalisation of each component exemplified in s1, referencing their corresponding definitions. Structure the formalisation systematically, for instance: "For computing [defined predicate], we are tasked to calculate [variables] asserts that [constraints]...". Refer to this step as Formalisation (s2) 3) Please consider the formalisation in s2 in detail, ensure this is correct and determine the entailment by breaking down the steps operating a symbolic representation. Combine variables, constants, and logical rules systematically at each step to find the solution. For clarity, provide clear reasoning for each step. Structure the explanation systematically, for instance: "Step 1: Calculate... Step 2:....". Refer to this step as Explanation (s3) 4) In conclusion, behind explaining the steps supporting the final answer to facilitate the final evaluation, extract the answer in a short and concise format by marking it as "output: entailment/contradiction" At this stage be strict and concise and refer to this step as Answering (s4). **Statement:** {statement} **Premise:** {premise} **Abstraction (s1):**

Table 9: SSR prompt for NLI4CT

tion of clinical reasoning. It comprises 2,400 premise–statement pairs, split 70/20/10 across train, test, and development sets. Examples where manually curated and labeled for entailment, with CTR sections and labels evenly distributed across all splits. Test set instances are annotated with 1-4 reasoning types, with an average of 1.6 types per instance. The distribution of reasoning types is shown in Table 12.

REACT Prompt

```
You are given a premise and a statement. Your task is to
determine the relationship between the statement and the
premise. Using thought, observation and action steps.
Instructions: - Begin with a Thought, where you explain
your reasoning step by step. - Use Action to simulate
operations like "Check if X is mentioned in the premise"
or "Compare X with Y". - Follow with Observation to
simulate what you learned from that action. - Repeat
this process until you reach a conclusion. - You must
end your response with "output: entailment" or "output:
contradiction".
Statement: {statement}
Premise: {premise}
Reasoning:
```

Table 10: REACT Prompt for NLI4CT

Self-Critique Prompt

```
You are given a premise and a statement. Your task is to
determine the relationship between the statement and the
premise. First generate a Draft Response, then generate
Critical Comments, then generate a Final Response.
Requirements: 1. Draft Response: Generate an initial
response 2. Critical Comments: Analyze your draft
response by considering: - Potential weaknesses or gaps -
Logical flaws or inconsistencies - Missing perspectives
or alternatives - Areas for improvement - Suggestions
for a better version - Steering toward the given answer
The critical comments should: - Be specific and
actionable - Reference particular parts of the draft -
Suggest concrete improvements - Consider different
angles or approaches - Guide towards a more
comprehensive solution
3. Final Response: Generate a final response that
incorporates the critical comments.
You must end your response with "output: entailment" or
"output: contradiction".
Statement: {statement}
Premise: {premise}
Draft Response:
```

Table 11: ISRR prompt for NLI4CT

Reasoning Type	Number of Instances
Lexical Equivalence	223
Clinical	131
Quantitative Comparison	148
Quantitative Derivation	95
Evidence	121
World-Knowledge Inference	84

Table 12: Number of instances requiring each type of reasoning in NLI4CT.

A.5 Generalisation Dataset Descriptions

A.5.1 MedNLI

MedNLI (Romanov and Shivade, 2018) is a CTNLI dataset constructed from de-identified patient notes in the MIMIC-III database (Johnson et al., 2016). Each instance pairs a short premise—typically a single sentence from a clinical note—with a hypothesis written by a medical expert and labeled as *Entailment*, *Contradiction*, or *Neutral*. For example, the premise *"He denied headache or nausea or vomiting"* and hypothesis *"He is afebrile"* is labeled *Neutral*, since the temperature status is not entailed or contradicted by the premise. For evaluation, we sample a balanced subset of 400 examples from the 1,422-instance test set.

A.5.2 TREC 2022 Clinical Trials Track

The TREC 2022 Clinical Trials Track (Roberts et al., 2022) presents a patient-to-trial retrieval task. Each input is a synthetic patient case description (topic) typically 5–10 sentences in length—that simulates an admission statement in an electronic health record. The retrieval corpus consists of approximately 375k CTRs from `ClinicalTrials.gov`, of which a subset has been labeled as *Eligible*, *Excluded*, or *Not Relevant*, relative to a given topic.

For this study, the TREC task is reformulated as a binary NLI classification problem. A balanced dataset of 400 trial–topic pairs is constructed, evenly split between the *Eligible* and *Excluded* classes. The *Not Relevant* category is omitted, as it primarily reflects retrieval failure rather than contradiction with eligibility criteria. An example instance is shown in Table 13.

Field	Content
Topic	*23-year-old man with exertional syncope, family history of sudden death, harsh systolic murmur, and septal hypertrophy.*
Trial Description	*Diagnostic strategies for suspected pulmonary embolism in outpatients.*
Label	Excluded

Table 13: Example instance from the TREC dataset.

A.6 Annotation Reliability Study

To assess the reliability of our expert reasoning-type annotations, we conducted an inter-annotator agreement (IAA) study on a random sample of 50 NLI4CT test instances, stratified to ensure coverage across the six reasoning types used in our analysis (Base entailment labels are provided by NLI4CT and were validated in Jullien et al. (2023a)) Each instance was independently annotated by three volunteer domain experts:

- one expert in natural language inference,
- one medical doctor,
- one physician associate.

Annotators were provided with the premise and statement, together with the definitions of the rea-

IAA Item ID	`8f078a17-14cd-4bbc-a9b6-b377ffa077b5`
Dataset / Split	NLI4CT / Test
Trial ID	NCT00143390
Section	Adverse Events
Gold NLI label	Contradiction

Instructions (Primary-type protocol). Read the premise and hypothesis. Select the *single* primary reasoning type needed to decide the NLI label. If multiple types apply, choose the best single category.

Premise .

Adverse Events 1: Total: 19/149 (12.75%); Anaemia 0/149 (0.00%); Acute myocardial infarction 1/149 (0.67%); Pericardial effusion 1/149 (0.67%); Prinzmetal angina 1/149 (0.67%); Meniere's disease 0/149 (0.00%); Vertigo 0/149 (0.00%); Cataract 2/149 (1.34%); Colitis ischaemic 1/149 (0.67%); Nausea 0/149 (0.00%); Vomiting 0/149 (0.00%); Chest pain 1/149 (0.67%).

Adverse Events 2: Total: 19/149 (12.75%); Anaemia 1/149 (0.67%); Acute myocardial infarction 0/149 (0.00%); Pericardial effusion 0/149 (0.00%); Prinzmetal angina 0/149 (0.00%); Meniere's disease 1/149 (0.67%); Vertigo 2/149 (1.34%); Cataract 1/149 (0.67%); Colitis ischaemic 0/149 (0.00%); Nausea 1/149 (0.67%); Vomiting 3/149 (2.01%); Chest pain 0/149 (0.00%).

Hypothesis. There was at least 1 recorded gastro-intestinal adverse event and 2 or more psychiatric events in the primary trial.

Primary reasoning type (select one)

- Clinical
- Lexical Equivalence
- Evidence
- World Knowledge
- Quantitative Comparison
- Quantitative Derivation

Figure 7: Example inter-annotator agreement (IAA) annotation form for a single NLI4CT instance under a primary-type reasoning protocol.

soning categories, and were asked to assign a *primary* reasoning type from {*Clinical*, *Lexical Equivalence*, *Evidence*, *World Knowledge*, *Quantitative Comparison*, *Quantitative Derivation*}. We use a primary-type protocol to support standard agreement measurement.

Agreement was measured using Fleiss' κ. Primary reasoning type agreement was $\kappa = 0.68$, indicating substantial agreement given the six-way categorisation and the known overlap between knowledge-driven categories (e.g., *Clinical* vs. *World Knowledge*). Example annotation form provided in Figure 7.

A.7 Bootstrap Confidence Intervals

For Quantitative Derivation ($n = 95$), we report 95% confidence intervals for the main category-level and prompt-level comparisons, computed using 1000 non-parametric bootstrap replicates with replacement and percentile-based bounds.

A.8 Results Tables

Model	F_1	Recall	Precision	Accuracy	% Valid
NLR					
gpt-4o-mini	0.692	0.700	0.722	0.699	1.000
llama 3.2 3B	0.536	0.843	0.551	0.573	0.978
DeepSeek-R1-Distill-Qwen-1.5B	0.559	0.747	0.556	0.573	0.925
Phi-4-mini-reasoning-3.8B	0.632	0.588	0.640	0.633	0.980
Qwen2.5-3B-Instruct	0.616	0.667	0.610	0.618	0.968
Qwen2.5-3B-Instruct LoRa	0.628	0.725	0.612	0.632	0.998
llama 3.2 3B LoRa	0.624	0.665	0.616	0.625	1.000
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.563	0.775	0.560	0.580	0.995
Phi-4-mini-reasoning-3.8B LoRa	0.626	0.505	0.678	0.632	0.958
SSR					
gpt-4o-mini	0.690	0.693	0.700	0.693	1.000
llama 3.2 3B	0.539	0.650	0.537	0.544	0.882
DeepSeek-R1-Distill-Qwen-1.5B	0.562	0.794	0.575	0.587	0.792
Phi-4-mini-reasoning-3.8B	0.560	0.497	0.573	0.561	0.980
Qwen2.5-3B-Instruct	0.535	0.553	0.560	0.536	0.765
Qwen2.5-3B-Instruct LoRa	0.579	0.599	0.576	0.579	0.985
llama 3.2 3B LoRa	0.614	0.619	0.613	0.614	0.945
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.503	0.697	0.516	0.519	0.920
Phi-4-mini-reasoning-3.8B LoRa	0.595	0.464	0.634	0.603	0.907
TAR					
gpt-4o-mini	0.642	0.655	0.681	0.655	1.000
llama 3.2 3B	0.606	0.775	0.585	0.616	0.963
DeepSeek-R1-Distill-Qwen-1.5B	0.536	0.674	0.540	0.545	0.895
Phi-4-mini-reasoning-3.8B	0.633	0.574	0.659	0.634	0.963
Qwen2.5-3B-Instruct	0.616	0.591	0.618	0.616	0.938
Qwen2.5-3B-Instruct LoRa	0.631	0.683	0.621	0.632	0.993
llama 3.2 3B LoRa	0.622	0.630	0.621	0.623	1.000
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.544	0.692	0.539	0.553	0.995
Phi-4-mini-reasoning-3.8B LoRa	0.635	0.635	0.635	0.635	0.988
ISRR					
gpt-4o-mini	0.632	0.657	0.717	0.657	1.000
llama 3.2 3B	0.529	0.530	0.530	0.529	0.988
DeepSeek-R1-Distill-Qwen-1.5B	0.450	0.716	0.498	0.484	0.790
Phi-4-mini-reasoning-3.8B	0.600	0.490	0.638	0.605	0.988
Qwen2.5-3B-Instruct	0.540	0.475	0.549	0.542	1.000
Qwen2.5-3B-Instruct LoRa	0.557	0.523	0.562	0.558	0.995
llama 3.2 3B LoRa	0.589	0.505	0.612	0.593	1.000
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.491	0.731	0.512	0.515	0.980
Phi-4-mini-reasoning-3.8B LoRa	0.579	0.360	0.706	0.604	0.998

Table 14: Results on TREC test set. Macro F_1, Precision and Recall.

Model	F_1	Recall	Precision	Accuracy	% Valid
NLR					
gpt-4o-mini	0.799	0.807	0.797	0.797	1.000
llama 3.2 3B	0.427	0.432	0.435	0.432	1.000
Qwen2.5-3B-Instruct	0.498	0.495	0.520	0.496	0.978
DeepSeek-R1-Distill-Qwen-1.5B	0.521	0.523	0.540	0.522	0.983
Phi-4-mini-reasoning-3.8B	0.690	0.691	0.693	0.691	0.995
llama 3.2 3B LoRa	0.624	0.645	0.664	0.645	1.000
Qwen2.5-3B-Instruct LoRa	0.652	0.658	0.653	0.657	1.000
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.526	0.530	0.529	0.530	1.000
Phi-4-mini-reasoning-3.8B LoRa	0.657	0.658	0.665	0.657	1.000
SSR					
gpt-4o-mini	0.799	0.800	0.800	0.800	1.000
llama 3.2 3B	0.413	0.458	0.517	0.461	0.960
Qwen2.5-3B-Instruct	0.630	0.626	0.653	0.626	0.995
DeepSeek-R1-Distill-Qwen-1.5B	0.420	0.423	0.430	0.424	0.790
Phi-4-mini-reasoning-3.8B	0.582	0.585	0.586	0.585	0.988
llama 3.2 3B LoRa	0.424	0.522	0.487	0.522	1.000
Qwen2.5-3B-Instruct LoRa	0.639	0.641	0.649	0.643	0.980
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.447	0.469	0.463	0.471	0.988
Phi-4-mini-reasoning-3.8B LoRa	0.640	0.647	0.665	0.648	0.953
TAR					
gpt-4o-mini	0.758	0.759	0.762	0.762	1.000
llama 3.2 3B	0.466	0.482	0.504	0.482	0.960
Qwen2.5-3B-Instruct	0.563	0.567	0.611	0.567	0.983
DeepSeek-R1-Distill-Qwen-1.5B	0.373	0.401	0.396	0.402	0.870
Phi-4-mini-reasoning-3.8B	0.574	0.574	0.592	0.573	0.978
llama 3.2 3B LoRa	0.465	0.555	0.561	0.555	1.000
Qwen2.5-3B-Instruct LoRa	0.541	0.575	0.578	0.575	1.000
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.368	0.420	0.425	0.420	1.000
Phi-4-mini-reasoning-3.8B LoRa	0.589	0.612	0.649	0.613	1.000
ISRR					
gpt-4o-mini	0.789	0.792	0.790	0.790	1.000
llama 3.2 3B	0.379	0.410	0.506	0.411	0.998
Qwen2.5-3B-Instruct	0.587	0.594	0.634	0.596	0.965
DeepSeek-R1-Distill-Qwen-1.5B	0.360	0.384	0.384	0.383	0.685
Phi-4-mini-reasoning-3.8B	0.577	0.575	0.594	0.575	0.988
llama 3.2 3B LoRa	0.439	0.519	0.512	0.521	0.993
Qwen2.5-3B-Instruct LoRa	0.579	0.585	0.601	0.587	0.945
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.395	0.428	0.461	0.432	0.902
Phi-4-mini-reasoning-3.8B LoRa	0.596	0.618	0.627	0.618	0.995

Table 15: Results on MedNLI test set. Macro F_1, Macro Precision, and Macro Recall

Model	F_1	Recall	Precision	Accuracy	% Valid
Direct					
gpt-4o-mini	0.643	0.696	0.630	0.644	1.000
llama 3.2 3B	0.490	0.817	0.521	0.532	0.962
Qwen2.5-3B-Instruct	0.501	0.327	0.533	0.519	0.990
DeepSeek-R1-Distill-Qwen-1.5B	0.560	0.414	0.601	0.571	0.978
Phi-4-mini-reasoning-3.8B	0.589	0.384	0.701	0.610	1.000
llama 3.2 3B LoRa	0.490	0.696	0.506	0.508	1.000
Qwen2.5-3B-Instruct LoRa	0.521	0.424	0.537	0.525	0.944
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.502	0.318	0.545	0.523	0.972
Phi-4-mini-reasoning-3.8B LoRa	0.542	0.313	0.650	0.573	0.998
NLR					
gpt-4o-mini	0.800	0.815	0.776	0.800	1.000
llama 3.2 3B	0.545	0.721	0.542	0.557	0.980
Qwen2.5-3B-Instruct	0.575	0.612	0.571	0.576	1.000
DeepSeek-R1-Distill-Qwen-1.5B	0.562	0.431	0.599	0.570	0.954
Phi-4-mini-reasoning-3.8B	0.680	0.645	0.693	0.680	0.994
llama 3.2 3B LoRa	0.701	0.660	0.721	0.702	1.000
Qwen2.5-3B-Instruct LoRa	0.631	0.540	0.665	0.634	1.000
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.647	0.552	0.687	0.650	1.000
Phi-4-mini-reasoning-3.8B LoRa	0.729	0.687	0.750	0.729	0.998
SSR					
gpt-4o-mini	0.759	0.727	0.832	0.760	1.000
llama 3.2 3B	0.441	0.819	0.492	0.495	0.808
Qwen2.5-3B-Instruct	0.568	0.708	0.556	0.575	0.884
DeepSeek-R1-Distill-Qwen-1.5B	0.479	0.660	0.494	0.493	0.754
Phi-4-mini-reasoning-3.8B	0.598	0.776	0.578	0.608	0.776
llama 3.2 3B LoRa	0.638	0.606	0.651	0.638	0.978
Qwen2.5-3B-Instruct LoRa	0.666	0.598	0.696	0.667	0.992
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.594	0.619	0.588	0.594	0.980
Phi-4-mini-reasoning-3.8B LoRa	0.662	0.710	0.647	0.663	0.996
TAR					
gpt-4o-mini	0.798	0.794	0.804	0.798	1.000
llama 3.2 3B	0.541	0.753	0.545	0.559	0.980
Qwen2.5-3B-Instruct	0.580	0.468	0.613	0.586	1.000
DeepSeek-R1-Distill-Qwen-1.5B	0.528	0.443	0.527	0.532	0.894
Phi-4-mini-reasoning-3.8B	0.686	0.669	0.689	0.686	0.976
llama 3.2 3B LoRa	0.682	0.684	0.681	0.682	1.000
Qwen2.5-3B-Instruct LoRa	0.667	0.663	0.668	0.667	0.998
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.584	0.560	0.586	0.584	0.996
Phi-4-mini-reasoning-3.8B LoRa	0.719	0.668	0.746	0.719	0.998
ISRR					
gpt-4o-mini	0.780	0.890	0.648	0.784	1.000
llama 3.2 3B	0.475	0.806	0.510	0.518	0.996
Qwen2.5-3B-Instruct	0.511	0.688	0.518	0.524	1.000
DeepSeek-R1-Distill-Qwen-1.5B	0.508	0.744	0.520	0.530	0.706
Phi-4-mini-reasoning-3.8B	0.672	0.690	0.665	0.672	0.994
llama 3.2 3B LoRa	0.674	0.574	0.722	0.677	0.998
Qwen2.5-3B-Instruct LoRa	0.641	0.466	0.744	0.653	0.998
DeepSeek-R1-Distill-Qwen-1.5B LoRa	0.552	0.348	0.644	0.575	0.992
Phi-4-mini-reasoning-3.8B LoRa	0.700	0.584	0.768	0.704	1.000

Table 16: Results on the NLI4CT test set. Macro F_1, Precision, and Recall are reported with entailment as the positive label.

Class	Base SSR	LoRa SSR	Base NLR	LoRa NLR	Base TAR	LoRa TAR	Base ISRR	LoRa ISRR	Base ZS	LoRa ZS
Quantitative Comp	0.428	0.652	0.565	0.789	0.499	0.723	0.461	0.727	0.508	0.455
QuantD	0.481	0.620	0.510	0.616	0.533	0.578	0.521	0.534	0.542	0.539
Lexical Equivalence	0.403	0.665	0.605	0.690	0.566	0.652	0.440	0.671	0.493	0.532
Clinical	0.463	0.592	0.478	0.677	0.497	0.668	0.417	0.667	0.442	0.477
World-Knowledge	0.394	0.617	0.583	0.688	0.529	0.679	0.466	0.625	0.453	0.503
Evidence	0.421	0.658	0.525	0.734	0.625	0.702	0.474	0.731	0.507	0.429

Table 17: Combined F1 scores for each class across all experimental settings with Llama-3.2-3B-Instruct on the NLI4CT test set. ZS = Zero-Shot.

Class	Base SSR	LoRa SSR	Base NLR	LoRa NLR	Base TAR	LoRa TAR	Base ISRR	LoRa ISRR	Base ZS	LoRa ZS
Quantitative Comp	0.639	0.681	0.594	0.694	0.650	0.716	0.520	0.702	0.501	0.582
QuantD	0.522	0.684	0.547	0.543	0.485	0.627	0.516	0.561	0.470	0.477
Lexical Equivalence	0.564	0.660	0.563	0.637	0.542	0.643	0.487	0.635	0.502	0.497
Clinical	0.508	0.648	0.588	0.631	0.618	0.643	0.493	0.626	0.527	0.510
World-Knowledge	0.513	0.636	0.595	0.660	0.646	0.724	0.616	0.611	0.386	0.575
Evidence	0.483	0.631	0.537	0.637	0.569	0.744	0.501	0.689	0.520	0.534

Table 18: Combined F1 scores for each class across all experimental settings with Qwen2.5-3B-Instruct on the NLI4CT test set. ZS = Zero-Shot.

Class	Base SSR	LoRa SSR	Base NLR	LoRa NLR	Base TAR	LoRa TAR	Base ISRR	LoRa ISRR	Base ZS	LoRa ZS
Quantitative Comp	0.454	0.672	0.660	0.637	0.547	0.590	0.567	0.573	0.631	0.541
QuantD	0.486	0.570	0.483	0.608	0.483	0.578	0.450	0.526	0.533	0.390
Lexical Equivalence	0.454	0.524	0.527	0.627	0.531	0.596	0.541	0.523	0.557	0.448
Clinical	0.433	0.594	0.615	0.621	0.550	0.563	0.451	0.500	0.558	0.563
World-Knowledge	0.613	0.566	0.463	0.578	0.470	0.482	0.533	0.521	0.565	0.537
Evidence	0.472	0.611	0.533	0.680	0.490	0.645	0.410	0.631	0.550	0.539

Table 19: Combined F1 scores for each class across all experimental settings with DeepSeek-R1-Distill-Qwen-1.5B on the NLI4CT test set. ZS = Zero-Shot.

Class	Base SSR	LoRa SSR	Base NLR	LoRa NLR	Base TAR	LoRa TAR	Base ISRR	LoRa ISRR	Base ZS	LoRa ZS
Quantitative Comp	0.600	0.668	0.696	0.736	0.712	0.775	0.685	0.749	0.598	0.591
QuantD	0.723	0.638	0.669	0.662	0.628	0.649	0.702	0.681	0.548	0.496
Lexical Equivalence	0.556	0.650	0.656	0.725	0.684	0.717	0.627	0.695	0.619	0.488
Clinical	0.579	0.626	0.694	0.695	0.662	0.661	0.668	0.677	0.608	0.634
World-Knowledge	0.574	0.701	0.713	0.713	0.648	0.676	0.602	0.693	0.545	0.454
Evidence	0.579	0.636	0.651	0.772	0.739	0.759	0.677	0.737	0.593	0.554

Table 20: Combined F1 scores for each class across all experimental settings with Phi-4-mini-reasoning on the NLI4CT test set. ZS = Zero-Shot.

Class	ISRR	TAR	NLR	SSR	ZS
Quantitative Comp	0.841	0.824	0.824	0.797	0.634
QuantD	0.652	0.768	0.810	0.767	0.623
Lexical Equivalence	0.798	0.768	0.778	0.711	0.619
Clinical	0.751	0.781	0.788	0.771	0.634
World-Knowledge	0.778	0.785	0.772	0.762	0.643
Evidence	0.805	0.868	0.809	0.727	0.618

Table 21: Combined F1 scores for each class across all experimental settings with GPT-4o mini on the NLI4CT test set. ZS = Zero-Shot.

Reasoning Type	Factor	Partial η^2	p-value
Clinical	Prompt	0.332	0.0118
Clinical	LoRA	0.289	0.0012
Clinical	Model	0.489	<0.001
World-Knowledge	Prompt	0.329	0.0125
World-Knowledge	LoRA	0.186	0.0122
World-Knowledge	Model	0.314	0.0078
Evidence	Prompt	0.435	0.0011
Evidence	LoRA	0.467	<0.001
Evidence	Model	0.395	0.0012
Quantitative Comp	Prompt	0.333	0.0117
Quantitative Comp	LoRA	0.293	0.0011
Quantitative Comp	Model	0.331	0.0054
QuantD	Prompt	0.300	0.0223
QuantD	LoRA	0.067	0.1472
QuantD	Model	0.526	<0.001
Lexical Equivalence	Prompt	0.390	0.0033
Lexical Equivalence	LoRA	0.261	0.0024
Lexical Equivalence	Model	0.395	0.0013

Table 22: Partial η^2 and p-values from fixed-effects Type II ANOVA models fitted separately for each reasoning type, with prompt, LoRA status, and model identity as predictors. For each reasoning type (Clinical, Common Sense, Existence, Numerical Comparison), we fitted a separate ordinary least squares (OLS) regression model predicting macro-F1 from three categorical predictors: prompt strategy, LoRA adaptation status, and model identity. We conducted a Type II ANOVA on each model to isolate the marginal effect of each factor, and computed partial eta-squared (η^2partial) to estimate the proportion of residual-adjusted variance attributable to each. This decomposition quantifies the unique contribution of each variable to performance variation within individual reasoning subdomains.

Class	Best Method	Best F_1	Margin
Clinical	NLR	0.625	0.0546
World-Knowledge	NLR	0.624	0.0530
Evidence	TAR	0.668	0.0855
Quantitative Comp	NLR	0.671	0.0617
QuantD	TAR	0.593	0.0351
Lexical Equivalence	NLR	0.629	0.0583

Table 23: The prompt with the highest F_1 score for each reasoning class, on the NLI4CT test set, and the margin over the average F_1 of other prompts.

Prompt	Avg. Number of Demonstrations (k)
NLR	1.52
SSR	1.46
ISRR	1.46
TAR	1.55
ZC	1.20

Table 24: Number of demonstrations per prompt variant, in thousands.

Factor	Eta^2	p-value	Effect Size
Prompt	0.440	0.001	Large
LoRA	0.357	0.0002	Large
Model	0.507	0.00006	Very Large

Table 25: Overall effect of prompt, LoRA, and model on F_1 performance across all samples (averaged across reasoning types). To assess the global contribution of prompt strategy, LoRA adaptation, and model architecture to performance variation we then fit an OLS model to F_1 score. with all predictors treated as fixed categorical effects. A Type II ANOVA was used to partition variance, and partial eta-squared values were calculated to quantify each factor's unique explanatory power. This analysis estimates how much each design choice contributes to overall model performance across tasks.

Reasoning Class	Mean F_1 (Lift vs. Base)
Clinical	0.614 (0.065)
World-Knowledge Inference	0.612 (0.060)
Evidence	0.653 (0.106)
Quantitative Comp	0.663 (0.081)
QuantD	0.579 (0.028)
Lexical Equivalence	0.614 (0.063)

Table 26: Average F_1 scores of LoRa models by reasoning class, with lift compared to corresponding base models in parentheses.

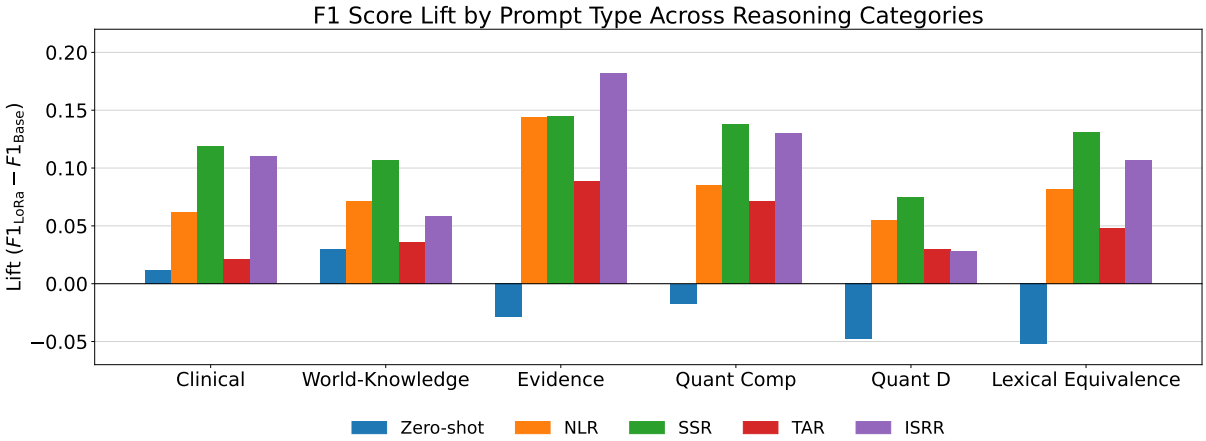

Figure 8: Impact of LoRA Tuning on F_1 Score by Prompt Strategy and Reasoning Type

Method	Class	LoRa Mean F_1	Base Mean F_1	Lift
NLR	Clinical	0.656	0.594	0.062
NLR	World-Knowledge Inference	0.660	0.589	0.071
NLR	Evidence	0.706	0.562	0.144
NLR	Quantitative Comp	0.714	0.629	0.085
NLR	QuantD	0.607	0.552	0.055
NLR	Lexical Equivalence	0.670	0.588	0.082
SSR	Clinical	0.615	0.496	0.119
SSR	World-Knowledge Inference	0.630	0.524	0.107
SSR	Evidence	0.634	0.489	0.145
SSR	Quantitative Comp	0.668	0.530	0.138
SSR	QuantD	0.628	0.553	0.075
SSR	Lexical Equivalence	0.625	0.494	0.131
TAR	Clinical	0.634	0.613	0.021
TAR	World-Knowledge Inference	0.640	0.604	0.036
TAR	Evidence	0.713	0.623	0.089
TAR	Quantitative Comp	0.701	0.630	0.071
TAR	QuantD	0.608	0.578	0.030
TAR	Lexical Equivalence	0.652	0.604	0.048
ISRR	Clinical	0.618	0.507	0.110
ISRR	World-Knowledge Inference	0.613	0.554	0.058
ISRR	Evidence	0.697	0.516	0.182
ISRR	Quantitative Comp	0.688	0.558	0.130
ISRR	QuantD	0.576	0.547	0.028
ISRR	Lexical Equivalence	0.631	0.524	0.107
Zero-shot	Clinical	0.546	0.534	0.012
Zero-shot	World-Knowledge Inference	0.517	0.487	0.030
Zero-shot	Evidence	0.514	0.543	-0.029
Zero-shot	Quantitative Comp	0.542	0.560	-0.017
Zero-shot	QuantD	0.476	0.523	-0.048
Zero-shot	Lexical Equivalence	0.491	0.543	-0.052

Table 27: Comparison of LoRa vs. base model performance (mean F_1) across reasoning classes and prompt methods on the NLI4CT test set. The "Lift" column shows the performance gain of the LoRa-tuned model over its base counterpart.

A.9 Related Work Table

Study	Reasoning	Prompt strategies
(Wen et al., 2025)	Math, Commonsense	Monologue, Self-ask, Self-debate
(Zhang et al., 2022)	Math, Commonsense, Symbolic	Manual vs. Auto CoT
(Press et al., 2022)	Multi-hop, Compositional	Self-ask vs. baseline
(Mondorf and Plank, 2024)	Math, Logical, Causal, Commonsense, Scientific, Social	CoT
(Khalid et al. 2025)	Relational (spatial, temporal)	Zero-shot, CoT
(Yu et al., 2025)	Mathematical	CoT

Table 28: Cross-domain studies contrasting prompting strategies and reasoning types.