Negation in Reasoning Traces: Interpretable Signals of Correctness and Provenance

Leon Hammerla and Alexander Mehler
{hammerla · mehler}@em.uni-frankfurt.de
Goethe University, Frankfurt am Main, Germany

Abstract

Chain-of-thought (CoT) reasoning is widely used in large language models (LLMs), but the resulting reasoning traces remain underexplored. We study these traces through the lens of discourse-level negation. Specifically, we distinguish between *corrective* negation, which rejects a prior reasoning step, and *refining* negation, which narrows or qualifies it, and introduce metrics to quantify their use in human- and LLM-authored reasoning traces. Across multiple benchmarks, we find that negation occurs much more frequently in intermediate reasoning traces than in final response texts. We then test whether negation-based features provide predictive and descriptive signal for correctness, model identity, and human-vs.-LLM authorship. For correctness prediction, negation-based features consistently outperform simple structural baselines and in several settings add complementary signal to embedding-based representations, although embeddings remain stronger overall. In a controlled comparison on correct human and LLM traces from the same dataset, our strongest results arise in human-vs.-LLM classification, where negation features outperform both structural and embedding baselines. Overall, these findings position discourse-level negation as an interpretable feature for reasoning-trace analysis, with especially strong utility for provenance-related classification and modest but consistent value for correctness prediction.

1 Introduction

Chain-of-thought (CoT) has made intermediate reasoning traces a prominent object of study in large language models (LLMs) (Wei et al., 2022). CoT presents problem solving as a sequence of intermediate steps that move from an initial, possibly flawed hypothesis toward a more accurate conclusion. CoT's stepwise reasoning loosely resembles classical forms of human inquiry. For example, the Socratic method (Benson, 2011) deepens understanding through iterative questioning, while the Hegelian dialectic (Forster, 1993) advances by synthesizing opposing ideas into higher-order insights. These analogies show how structured reasoning sequences refine conclusions (Pei et al., 2025; Abdali et al., 2025), using negation (Schon et al., 2021) to signal correction, opposition, or refinement. In reasoning, negation may appear in at least two operations: (1) as a corrective operation that rejects a prior line of thought entirely, or (2) as a refining operation that narrows and sharpens reasoning without discarding it completely. We hypothesize that these two operations are salient discourse markers of stepwise reasoning and are reflected in both human- and model-generated reasoning traces. Modeling how LLMs employ negation during reasoning can provide insights into the dynamics of stepwise reasoning, as well as into variation in response correctness and trace style. However, while prior work on LLM reasoning traces has examined their structure (Bogdan et al., 2025), interpretability (Bhambri et al., 2025), and effectiveness (Gandhi et al., 2025), the role of negation within such traces remains largely unexplored, with few exceptions (Ye et al., 2023). We address this research gap by making three contributions:

1. We build a dataset of human and LLM reasoning traces, auto-labeled for two negation types and partially validated through human evaluation.

2. We introduce several metrics to quantify negation in human and LLM reasoning traces.

3. We use these metrics to examine how negation patterns in reasoning traces relate to response correctness, model identity, and human authorship.

Proceedings of the 6th Workshop on Natural Language Meets Logic and Machine Learning (NALOMA), pages 19–39

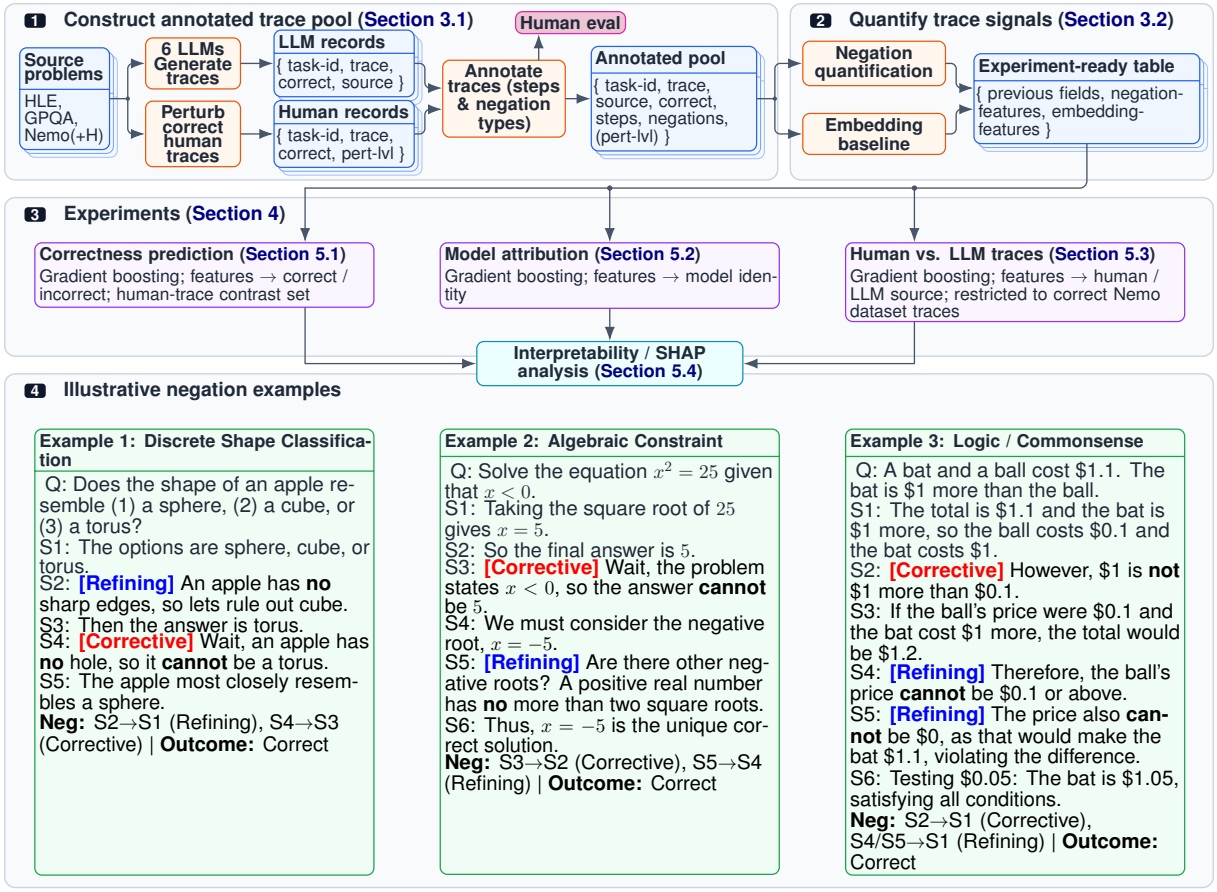


Figure 1: Pipeline for constructing our annotated reasoning-trace dataset, extracting negation-sensitive features, and evaluating the resulting representations across different tasks. Representative annotated traces illustrate our distinct negation types.

2 Related Work

2.1 Reasoning Traces in LLMs

Chain-of-thought (CoT) prompting (Wei et al., 2022) established that large language models can often improve performance on multi-step tasks when prompted to produce intermediate reasoning steps. Subsequent work expanded this general paradigm through methods such as self-consistency (Wang et al., 2023), Tree/Graph of Thoughts (Yao et al., 2023; Besta et al., 2024), and Program of Thoughts (Chen et al., 2023), which vary the structure or execution of intermediate reasoning. More recently, reasoning-oriented open-weight models such as Qwen3 (Yang et al., 2025), DeepSeek-R1 (Guo et al., 2025), and gpt-oss (OpenAI et al., 2025) have made such traces increasingly accessible through explicit "thinking" or reasoning modes.

2.2 Analyzing LLM Reasoning Traces

As reasoning traces have become more available, recent work has shifted attention from final-answer accuracy alone to the analysis of the traces themselves. Prior studies examine, among other things, which intermediate steps are most influential for correct answers (Bogdan et al., 2025; Zhang et al., 2025a), how traces align with structured knowledge sources (Zhang et al., 2025b), how interpretability relates to model performance (Bhambri et al., 2025), and how reasoning length affects controllability and effectiveness (Marjanovic et al., 2026). Other work studies reasoning chains as a route to model improvement (Gandhi et al., 2025), compares LLM reasoning with human cognition (Marjanovic et al., 2026), and questions the extent to which LLM traces reflect logical reasoning (Xu et al., 2025). Overall, this literature has shown that reasoning traces can be studied in terms of usefulness, step importance, faithfulness, interpretability, and error propagation, rather than only through final prediction accuracy.

2.3 Negation, Discourse, and Self-Revision in Reasoning

Our work is most closely related to research that treats reasoning traces as structured discourse rather than merely as sequences of tokens or steps. In linguistic and discourse-theoretic terms, negation plays a central role in rejection, correction, narrowing, and contrast (Horn, 1989; Fălăuş, 2020), and argumentative work has shown that it can function to counter or delimit prior commitments (Apothéloz et al., 1993). These functions are especially relevant to stepwise reasoning, where later steps often revise, restrict, or overturn earlier ones. In this sense, corrective and refining negation can be viewed as discourse-level markers of self-revision, complementing broader frameworks that analyze reasoning traces in terms of cognitive or functional elements (Kargupta et al., 2025). However, despite growing interest in LLM reasoning traces, prior work has rarely operationalized explicit discourse-level linguistic phenomena within them. Most existing studies focus on performance, length, structure, interpretability, or faithfulness, rather than on how particular linguistic devices organize revision across steps. Negation in particular has received little systematic attention as a trace-level feature, despite its natural connection to error correction, hypothesis narrowing, and self-revision during reasoning. We address this gap by modeling corrective and refining negation as interpretable discourse features and testing whether their distribution in reasoning traces carries signal about correctness, model identity, and human-versus-LLM authorship.

3 Methodology

3.1 Dataset

We construct our dataset from three sources[1]: three established benchmarks (GPQA Diamond (gpqa) (Rein et al., 2024), AIME 25 (aime25) (Zhang and Math-AI, 2025), and the Nemotron-Math-HumanReasoning (nemo) dataset (Du et al., 2025), which provides human-authored CoT traces for Olympiad-level mathematics problems from the AoPS forum (comparable in difficulty to AIME 25). In the first stage, we prompt six reasoning-capable LLMs to solve each benchmark task (full model list in Figure 2). The prompt instructs models to generate an explanation and return a JSON-formatted

[1]Our datasets are hosted on Hugging Face at https://huggingface.co/RT-NEG.

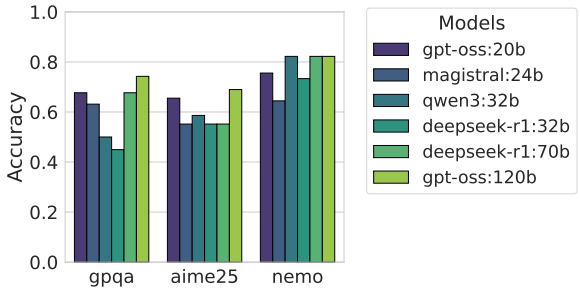


Figure 2: Benchmark results of all models evaluated on the gpqa, aime25, and the nemo dataset.

conclusion for unambiguous parsing (prompt template in Appendix F.1). The intermediate reasoning trace is implicitly generated by each model's native reasoning process. As the human dataset contains only correct solutions, we generate incorrect examples to balance prediction tasks. We use Llama3.3-70B as a meta-annotator to inject human-style reasoning errors at three severity levels, each yielding an incorrect conclusion. These corrupted traces enable us to construct both positive and negative human samples (prompt template in Appendix F.2). In the second stage, we annotate all LLM and human reasoning traces using Llama3.3-70B. Following (Bogdan et al., 2025), the annotator first segments each trace into reasoning steps, typically aligned with individual sentences. It then labels each step with one of two negation types using a dedicated prompt (see Appendix F.3). Building on linguistic theories of negation as a logic-based means to cancel alternatives (Horn, 1989; Fălăuş, 2020) and argumentative models showing its use in countering or delimiting conclusions (Apothéloz et al., 1993), we distinguish two pragmatic functions of explicit negation in reasoning: (1) **refining negation**, which rules out previously considered alternatives to narrow the solution space, and (2) **corrective negation**, which explicitly rejects or revises a prior step, sometimes without offering an alternative. Both types span reasoning steps rather than single sentences, reflecting how negation dynamically manages discourse commitments and alternatives (Büring, 2008; Krifka, 2003).

3.1.1 Preliminary Analysis

Before introducing our negation metrics, we conduct four preliminary checks to verify that reasoning traces are suitable for discourse-level negation analysis. First, across datasets, *gpt-oss* performs strongest overall, with *gpt-oss:120b* leading in most settings, while *DeepSeek-R1:70b* re-

mains competitive on *gpqa* and *nemo* but weaker on *aime25* (Figure 2). Second, using D-Neg (Hammerla et al., 2025), which builds on NegBERT (Khandelwal and Sawant, 2020), we find that reasoning traces contain substantially more negation than final responses in almost all dataset-model combinations, often by a factor of 2× to 30×. This indicates that negation is considerably more salient in intermediate reasoning than in answer-only text (Figure 3). Third, we perform a coarse sanity check on our distinction between refining and corrective negation. Across datasets, refining negations show a positive aggregate association with accuracy, whereas corrective negations show a negative one. We treat this only as supporting evidence that the two categories capture different aspects of reasoning behavior, rather than as a standalone predictive result.[2] Finally, to assess annotation reliability, two linguists independently reviewed 36 reasoning traces[3] sampled across datasets and models. Although this sample is limited at the trace level, it comprises 634 reasoning steps, each of which required a judgment about whether it contains discourse-relevant negation and, when applicable, which negation type it instantiates. Inter-annotator agreement between the two human annotators was moderate ($\kappa = 0.68$), whereas agreement between the LLM annotator and the human annotators was lower ($\kappa = 0.37$ and 0.38), as expected for a challenging discourse-level phenomenon (Mirzakhmedova et al., 2024; Li et al., 2025). We therefore treat the resulting labels as weak but operationally useful proxy annotations for aggregate analysis.

3.2 Metrics

We quantify negation to examine its role in reasoning, where it signals self-monitoring, error detection, and logical refinement (Horn, 1989). Refining negations often reflect hypothesis adjustment; corrective ones mark prior inconsistencies. Quantifying them reveals how reasoning unfolds and links to performance or reasoning specifics of LLMs.

3.2.1 Metric Design

To capture the complexity of negation use, we designed metrics across three key dimensions: **composition**, which captures the frequency, presence pattern, and type balance of negations; **position**, which captures when negations appear and thus the timing of revisions; and **structure**, which captures how negations cluster or persist across steps, indicating reflective consistency.

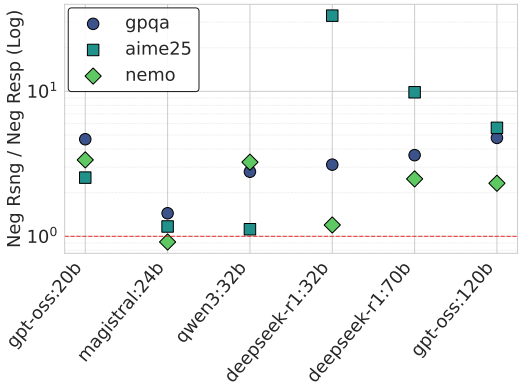


Figure 3: Ratio of average number of negations per 100 tokens in reasoning traces relative to responses across multiple datasets and models. A value greater than one (see red dashed line) indicates that negations occur more frequently in reasoning traces than in responses.

1. Compositional Trace Statistics This category captures the compositional properties of negations within a reasoning trace. Specifically, it indicates whether refining and/or corrective negations are present, and quantifies the frequency of each negation type relative to the total number of reasoning steps. It also includes a score for the balance between corrective and refining negations, providing insight into how evenly the two forms are represented within the trace:

- **Norm Neg:** Number of negations type t normalized by total number of reasoning steps:

$$\tilde{N}^t_{\text{neg}} = \frac{N^t_{\text{neg}}}{N_{\text{rstep}}}$$

  where $N_{\text{rstep}} = |T|$ is the total number of reasoning steps in reasoning trace T and $N^t_{\text{neg}} = |\text{Negs}^t|$ is the total number of corrective and refining negations in the reasoning trace ($t \in \{\text{cor}, \text{ref}\}$ is the negation type and Negs^t the set of all negations in T of type t).

- **RT Label:** Categorical label indicating which types of negation appear in the trace:

$$L = \begin{cases} 1, & N^{\text{cor}}_{\text{neg}} > 0 \text{ and } N^{\text{ref}}_{\text{neg}} = 0 \\ 2, & N^{\text{cor}}_{\text{neg}} = 0 \text{ and } N^{\text{ref}}_{\text{neg}} > 0 \\ 3, & N^{\text{cor}}_{\text{neg}} > 0 \text{ and } N^{\text{ref}}_{\text{neg}} > 0 \\ 0, & \text{otherwise} \end{cases}$$

[2] See Appendix B for full methodology, binning, and smoothed trends.

[3] Using the same guidelines as the LLM; see Appendix F.3.

- **Diversity Score:** Measures the balance of negation types in trace T:

$$D = \sum_{t \in \{\text{cor,ref}\}} p_t^2$$

where $p_t = \frac{N_{\text{neg}}^t}{N_{\text{neg}}^{\text{cor}} + N_{\text{neg}}^{\text{ref}}}$.

2. Positional Characteristics These metrics capture *where* negations occur within reasoning traces (e.g., average normalized position, first/last negation type, and frequency in the first vs. second half). They reveal whether negations tend to appear early, late, or evenly, reflecting the temporal dynamics of self-monitoring.

- **Norm Avg Pos:** Average position of each negation type t in the trace T, normalized by total number of reasoning steps:

$$\tilde{P}^t = \frac{1}{N_{\text{neg}}^t} \sum_{i=1}^{N_{\text{neg}}^t} \frac{\text{step\_idx}(\text{Negs}_i^t)}{N_{\text{rstep}}}$$

- **Edge Neg Type** Type of the first and last negation in the trace:

$$L_{\text{edge}} = \begin{cases} 1, & \text{ntype}(\text{Negs}_{\text{edge}}) = \text{cor} \\ 2, & \text{ntype}(\text{Negs}_{\text{edge}}) = \text{ref} \\ 0, & \text{otherwise} \end{cases}$$

where edge $\in \{\text{first}, \text{last}\}$ and Negs is the set of all negations in trace T.

- **Neg Frequency in first vs. second half:** Ratio of type-t negations in the first vs. second half of trace T:

$$F^t = \frac{|\text{Negs}_{\text{first half}}^t|}{|\text{Negs}_{\text{second half}}^t|}$$

3. Sequential Structure and Dependency This category captures structural patterns of negation across reasoning steps, including gap statistics (maximum span and variance of inter-negation distances), sequence patterns (number of type switches, longest consecutive runs), randomness tests, and autocorrelation. These metrics reveal whether negations are scattered or follow structured, self-regulatory sequences.

- **Max Neg Span:** Max number of steps between *gt*-type negations, $gt \in \{\text{cor}, \text{ref}, \text{any}\}$:

$$G_{\max}^{gt} = \max(\text{Gaps}^{gt})$$

- **Neg Gap Variance:** Variance of gaps between consecutive gt-type negations ($N^{gt} = |\text{Gaps}^{gt}|$):

$$\text{Var}[\text{Gaps}^{gt}] = \frac{1}{N^{gt}} \sum_{i=1}^{N^{gt}} \left(\text{Gap}_i^{gt} - \overline{\text{Gap}^{gt}}\right)^2$$

- **Neg Patterns:** Number of switches between corrective and refining negations, and longest consecutive sequence of each negation type.

$$S = |\text{C}_{\text{cor}\to\text{ref}}| + |\text{C}_{\text{ref}\to\text{cor}}|$$
$$L^t = \max(\hat{C}^t)$$

$\text{C}_{\text{cor}\to\text{ref}}$ and $\text{C}_{\text{ref}\to\text{cor}}$ are the sets of the corresponding type switches; $\hat{C}^t$ is the set of consecutive sequences of negations of type t.

- **Wald-Wolfowitz Test:** Tests randomness of the binary sequence representing the trace ($0 =$ no negation, $1 =$ negation of type t).

$$\mathbb{E}[R^t] = \frac{2n_1^t n_2^t}{n_1^t + n_2^t} + 1$$

$$\text{Var}[R^t] = \frac{2n_1^t n_2^t (2n_1^t n_2^t - n_1^t - n_2^t)}{(n_1^t + n_2^t)^2 (n_1^t + n_2^t - 1)}$$

n_1^t (n_2^t): number of steps with(out) negation of type t. The standardized test statistic is:

$$Z^t = \frac{R^t - \mathbb{E}[R^t]}{\sqrt{\text{Var}[R^t]}}$$

- **Autocor:** Autocorrelation for different lags ($k \in [1, .., 8]$) for negations of type t:

$$\rho_k^t = \frac{\sum_{i=1}^{N-k} (Y_i^t - \overline{Y^t})(Y_{i+k}^t - \overline{Y^t})}{\sum_{i=1}^{N} (Y_i^t - \overline{Y^t})^2}$$

- **Integrated Autocor:** Measures persistence of negations of type t across multiple steps:

$$\tau^t = 1 + 2 \sum_{k=1}^{\text{max\_lag}} \rho_k^t$$

3.2.2 Metric Complementarity

To assess redundancy in the proposed metric set, we compute pairwise correlations between all negation-based metrics. Figure 4 shows that most metrics are weakly correlated or uncorrelated, indicating that they capture distinct aspects of negation behavior. The main exceptions are two larger

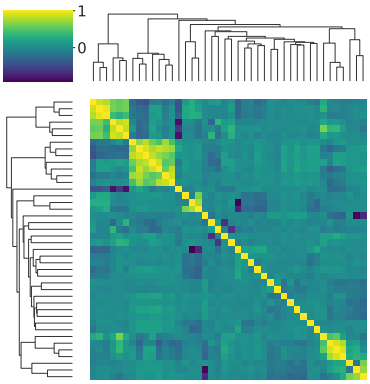


Figure 4: Hierarchically clustered heatmap of correlations between negation-quantifying metrics.

clusters corresponding primarily to autocorrelation features for refining and corrective negation across different lag sizes, where correlation is expected. Overall, the heatmap supports retaining the full metric set, as redundancy is limited and mostly confined to closely related temporal-dependency metrics.

4 Experiments

We investigate whether discourse-level negation patterns in reasoning traces provide information about (i) response correctness, (ii) model identity, and (iii) authorship, i.e. whether a trace was produced by a human or an LLM. Our experiments are organized accordingly: we first study correctness prediction, beginning with single-metric analyses and then moving to multi-feature classification, including an additional evaluation on perturbed human reasoning traces. We then examine model identity prediction and, finally, human-vs.-LLM authorship classification. Unless otherwise noted, all classification experiments use a Gradient Boosting (GB) LightGBM model with 1000 estimators, learning rate 0.01, maximum depth 3, minimum of 2 samples per leaf, subsample ratio 0.8, column subsampling of 0.8, and balanced class weights. All models are evaluated with five-fold cross-validation (CV), and we report fold-averaged results.[4] For each classification task, we compare four feature settings: (1) our negation-based feature set, (2) a structural baseline consisting of reasoning-trace length, total number of reasoning steps, and number of sentences, (3) an embedding baseline based on sentence-transformer representations from `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019), and (4) a combined representation that concatenates the negation features with the embedding features.

Correctness prediction. We begin with single-metric analyses to assess how individual negation-related metrics relate to response correctness. To capture both linear and non-linear effects, we report correlation, mutual information (MI), area under the ROC curve (AUC), and Cohen's d effect size for individual features. We then turn to multi-feature classification, where the goal is to predict whether a reasoning trace leads to a correct or incorrect final answer. These experiments are conducted across our benchmark datasets and model outputs using the four feature settings described above. For human reasoning traces, we additionally evaluate correctness prediction across the three perturbation levels introduced in Section 3.1. Incorrect traces are generated by perturbing correct human traces with increasing severity. This setup allows us to examine how classification performance changes as reasoning errors become more pronounced and correct and incorrect traces become more easily separable. Each perturbation-level dataset is balanced, containing equal numbers of correct and incorrect traces.

Model identity prediction. We next test whether reasoning traces contain model-specific stylistic signal by predicting which LLM produced a given trace. We use the same GB configuration and the same four feature settings as above in order to compare the discriminative value of negation patterns against structural and embedding-based baselines.

Human-vs.-LLM authorship classification. Finally, we evaluate whether reasoning traces authored by humans can be distinguished from those generated by LLMs. To this end, we construct a balanced dataset by pairing all human traces from the *nemo* dataset with an equal number of randomly selected correct LLM-generated traces drawn from the nemo dataset as well. We again compare negation features, the structural baseline, the embedding baseline, and the combined negation+embedding representation under the same evaluation setup.

[4] Experiments primarily use scikit-learn (Pedregosa et al., 2011) and LightGBM (Lamb et al., 2025).

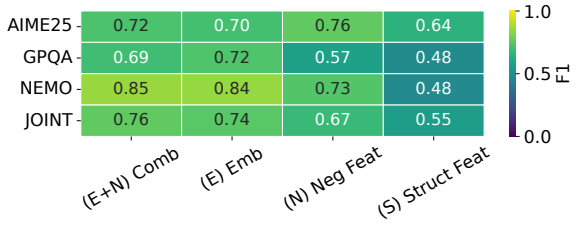


Figure 5: Five-fold CV binary F1 scores for correctness prediction across all datasets and the merged dataset across all different feature settings.

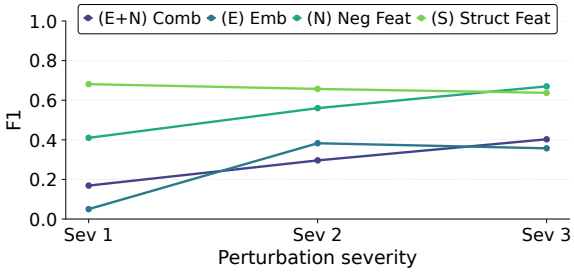


Figure 6: Five-fold CV binary F1 scores for classifier performance on human reasoning traces under varying levels of human-like perturbation severity.

5 Results and Analysis

5.1 Correctness Prediction

5.1.1 Single-Feature Analysis

Across all single-feature analyses, individual metrics exhibit only limited predictive power for distinguishing correct from incorrect reasoning traces. Absolute correlations remain low (max. ≈ 0.1), and mutual information scores peak at roughly 0.02. Cohen's d shows somewhat larger separations between the distributions, with several metrics reaching values around 0.2, corresponding to small to moderate effect sizes. Similarly, AUC values exceed the random baseline only marginally for a small subset of features. A consistent pattern emerges, however: metrics associated with refining negations dominate among the strongest-performing features across all measures (24 refining vs. 2 corrective). This suggests that refining-related behavior is more systematically linked to model correctness than corrective negation, reinforcing our observations from the preliminary analysis (Appendix C).

5.1.2 Multi-Feature Correctness Prediction

We next evaluate multi-feature correctness prediction with Gradient Boosting, reporting five-fold CV F1 scores for the four feature settings introduced in Section 4. Across the benchmark datasets and their joint aggregation, negation features consistently outperform the structural baseline, indicating that discourse-level negation captures signal beyond simple trace statistics such as length, number of steps, and sentence count. Averaged across datasets, this improvement amounts to 27.85%, with a large standardized effect size ($d = 2.03$). At the same time, embedding-based representations remain stronger overall: on *nemo*, *gpqa*, and the joint dataset, embedding features outperform negation features, whereas on *aime25*, negation features achieve a higher F1 score by 0.06. Overall, moving from negation features to embeddings yields an average relative improvement of 10.98% ($d = -0.74$), suggesting that embeddings provide a stronger standalone representation for correctness prediction. However, combining negation features with embeddings yields small additional gains in several settings, suggesting that negation captures some complementary information beyond generic embedding-based representations (Figure 5). Performance also varies across model families and datasets. No single model dominates in every condition: *magistral:24b* performs best on *aime25* (mean F1 0.643), *deepseek-r1:70b* on *gpqa* (0.753), *qwen3:32b* on *nemo* (0.900), and *gpt-oss:120b* on the joint dataset (0.737). Across all conditions, *gpt-oss:120b* achieves the highest average F1 score (0.719). Within both the *gpt-oss* and *deepseek-r1* families, larger models generally outperform smaller ones, although this trend is not uniform across all datasets (see Figure 11).

5.1.3 Human-Trace Perturbation Study

On the human reasoning traces from the *nemo* dataset, we conduct a controlled perturbation study in which incorrect traces are constructed by introducing human-like reasoning errors into originally correct human CoTs. We then repeat the multi-feature correctness experiment using these perturbed traces as negative samples and report fold-averaged binary F1 scores across the three perturbation severity levels (see Section 3.1). Because these negative samples are synthesized rather than naturally occurring, we interpret this setting as a controlled stress test of feature sensitivity to injected reasoning errors rather than as a direct model of authentic human failure. Across all three severity levels, the structural baseline performs unexpectedly strongly, achieving fold-averaged F1 scores above 0.6 throughout. At the same time,

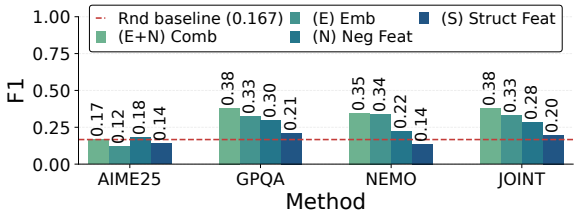


Figure 7: Five-fold CV macro F1 scores for model detection from reasoning traces grouped by datasets and feature settings.

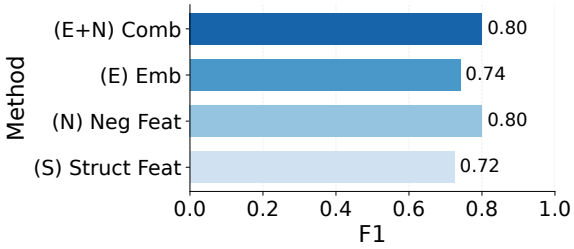


Figure 8: Five-fold CV binary F1 scores for human-vs.-LLM authorship classification with different feature sets.

its performance decreases as perturbation severity increases, whereas the negation-based representation becomes more discriminative as the injected reasoning errors grow more pronounced, reaching its best performance at the highest severity level. Embedding-based features perform substantially worse than both the structural and negation-based representations in this setting. One plausible explanation is that the perturbations preserve much of the lexical content of the original trace while altering its explicit discourse structure, making negation- and structure-based features relatively more informative than generic sentence embeddings. Overall, this controlled result suggests that negation-based features are sensitive to injected reasoning distortions in human-authored traces, although the finding should not be generalized directly to naturally occurring incorrect human reasoning.

5.2 Model Identity Prediction

Next, we test whether reasoning traces contain identifiable model-specific stylistic signatures. Concretely, we ask whether the originating LLM can be predicted from the four feature settings defined in Section 4. Because this is a multiclass classification task, we report five-fold CV macro F1 scores. Overall, model identity prediction proves challenging: with six models, the random baseline is 0.166 macro F1, and even the best results remain modest in absolute terms. Nevertheless, negation features consistently outperform the structural baseline across all four datasets (*aime25*, *nemo*, *gpqa*, and the joint dataset), indicating that they capture stylistic signal beyond simple trace statistics. Embedding-based features remain stronger overall, but combining embeddings with negation features yields further improvements in every setting, reaching a peak macro F1 of 0.38 on the joint dataset. Taken together, these results suggest that reasoning traces contain a limited but non-negligible amount of model-specific signal, and that negation contributes useful complementary information for capturing it (Figure 7).

Feature	Group	Human	LLM	SMD	q_{group}
`cor_neg_longest_run`	Struct.	1.111	0.306	0.561	$< .001$
`cor_negs_normed`	Comp.	0.125	0.051	0.623	.017
`ref_negs_normed`	Comp.	0.031	0.079	-0.476	.021

Table 1: Group-wise FDR-significant differences between human and LLM reasoning traces in the authorship classification dataset. Positive SMD values indicate higher values for human traces; negative values indicate higher values for LLM traces. Only features satisfying $q_{\text{group}} \leq .05$ and $|\text{SMD}| \geq .2$ are shown.

5.3 Human-vs.-LLM Authorship Classification

Finally, we evaluate whether human-authored reasoning traces can be distinguished from LLM-generated ones. As in the other binary classification settings, we report five-fold CV binary F1 scores. Unlike model identity prediction, this task exhibits a much stronger signal across all feature settings. Negation features perform best overall, reaching an average F1 score of 0.80 and outperforming both the embedding baseline (0.74) and the structural baseline (0.72). Combining negation features with embeddings does not yield further improvement, with the combined representation remaining at the same maximum F1 score of 0.80. These results suggest that human-vs.-LLM authorship is substantially easier to detect than model identity in reasoning traces, and that discourse-level negation features are particularly effective for capturing differences between human and model-generated reasoning traces (Figure 8).

5.4 Interpretability

To better understand which aspects of negation drive performance across tasks, we compute SHAP values for correctness prediction on the joint dataset, model identity prediction, and human-vs.-

LLM authorship classification. Detailed SHAP plots are provided in Appendix E. For interpretability, we aggregate features into broader classes and summarize their importance using mean absolute SHAP values normalized to sum to 1 within each task. A first clear pattern is that the relative importance of *refining* and *corrective* negation differs by task. For correctness prediction, refining features contribute more strongly than corrective ones (64% vs. 36%), consistent with both our preliminary analyses and the single-feature results. This suggests that metrics quantifying refining negation provide more information for correctness prediction than metrics quantifying overt corrective rejection of prior steps. Importantly, this does not imply that correct traces necessarily contain more refining negation, nor that refining negation is causally more effective. Rather, it indicates that variation in refining-related features is more useful for distinguishing correct from incorrect model reasoning traces in our classification setting. By contrast, contributions are more balanced for model identity (43% vs. 57%) and human-vs.-LLM authorship classification (58% vs. 42%), indicating that provenance-related tasks depend on a broader mix of negation behavior. A second pattern concerns feature type. Across all three tasks, structural features are the most consistently informative dimension (average 41%), ahead of positional (29%) and compositional features (30%). This suggests that *how* negation is organized across a reasoning trace is at least as important as *how much* negation occurs. Overall, the SHAP analysis supports the usefulness of the corrective/refining distinction and indicates that different downstream tasks rely on different aspects of negation behavior. To complement this model-based interpretation with a direct descriptive comparison, we compare human and LLM traces in the authorship setting across all negation features. For numeric features, we compute group means, standardized mean differences (SMD), and Mann-Whitney U tests; for categorical features, we use contingency-table statistics. Since the metrics were defined a priori as compositional, positional, and structural feature families, we apply Benjamini-Hochberg FDR correction within each family and report features as significant only if they satisfy $q \leq 0.05$ and a small effect-size threshold ($|\mathrm{SMD}| \geq 0.2$ for numeric features). Only three features pass these criteria (Table 1): humans show more corrective negation per reasoning step and longer repeated runs of corrective negations, whereas LLMs show more refining negation per reasoning step. No positional feature passes the group-wise FDR criterion, suggesting that the strongest human-LLM differences concern the function and sequential organization of negation rather than its location in the trace. These descriptive results clarify the SHAP analysis: the authorship signal is concentrated in a small set of interpretable negation behaviors, with human traces characterized more by corrective revision and LLM traces more by refining restriction.

6 Conclusion

We introduced a discourse-level approach to analyzing reasoning traces through negation. Specifically, we distinguished between *refining* and *corrective* negation, assembled a dataset of human- and LLM-authored reasoning traces annotated for these two functions, and proposed metrics that quantify negation in terms of composition, position, and sequential structure. Using these representations, we tested whether negation patterns carry signal for correctness, model identity, and human-vs.-LLM authorship. Across correctness experiments on LLM-generated traces, negation features consistently outperformed simple structural baselines, although embedding-based representations remained stronger overall and negation provided mainly complementary gains in combined settings. In a controlled perturbation study on human traces, negation features proved especially sensitive to injected reasoning distortions, though this result should be interpreted separately from naturally occurring human error. For model identity prediction, performance remained modest overall, but negation still contributed useful signal beyond structural cues. Our strongest results were obtained for human-vs.-LLM authorship classification on correct traces from the same benchmark, where negation features outperformed both structural and embedding baselines. Overall, these findings suggest that discourse-level negation is an interpretable and informative feature for reasoning-trace analysis, especially for provenance-related distinctions and, to a lesser extent, correctness-related variation. More broadly, the results motivate further work on reasoning traces through explicit linguistic phenomena, including revision, uncertainty, and inter-sentential propositional relations.

Limitations

This work is subject to several scope-related limitations:
(1) First, we focus on explicit instances of negation, distinguishing between refining and corrective uses within reasoning traces. While this provides a clear and interpretable operationalization, it does not capture more implicit or indirect forms of negation, such as pragmatic denial or contrastive framing, which may also influence reasoning.
(2) Second, our analysis is centered on CoT–style reasoning traces, which offer an explicit, step-wise view and naturally encode structured dependencies across steps. Other reasoning frameworks or prompting strategies may exhibit different surface realizations of negation, and extending the analysis to such settings remains a promising direction for future work.
(3) Finally, this study covers only a subset of task types that elicit explicit reasoning (restricted to textual, single-modality tasks). Although the datasets include both open-ended and multiple-choice problems, we hypothesize that negation usage patterns vary with task characteristics.
Overall, these considerations reflect design choices aimed at analytical clarity and point to natural extensions rather than fundamental shortcomings of the proposed approach.

Acknowledgements

This work was supported by the German Research Foundation (DFG) as part of Project INF within CRC 1629 "Negation in Language and Beyond" (NegLaB - https://www.neglab.de/).

References

Sara Abdali, Can Goksen, Michael Solodko, Saeed Amizadeh, Julie E. Maybee, and Kazuhito Koishida. 2025. Self-reflecting large language models: A hegelian dialectical approach. *Preprint*, arXiv:2501.14917.

Denis Apothéloz, Pierre-Yves Brandt, and Gustavo Quiroz. 1993. The function of negation in argumentation. *Journal of Pragmatics*, 19(1):23–38.

Hugh H Benson. 2011. Socratic method. *The Cambridge companion to socrates*, pages 179–200.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. 2025. Do cognitively interpretable reasoning traces improve llm performance? *Preprint*, arXiv:2508.16695.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which llm reasoning steps matter? *Preprint*, arXiv:2506.19143.

Daniel Büring. 2008. The least at least can do. In *West Coast Conference on Formal Linguistics (WCCFL)*, volume 26, pages 114–120.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Wei Du, Branislav Kisacanin, George Armstrong, Shubham Toshniwal, Ivan Moshkov, Alexan Ayrapetyan, Sadegh Mahdavi, Dan Zhao, Shizhe Diao, Dragan Masulovic, Marius Stanean, Advaith Avadhanam, Max Wang, Ashmit Dutta, Shitij Govil, Sri Yanamandara, Mihir Tandon, Sriram Ananthakrishnan, Vedant Rathi, and 6 others. 2025. The challenge of teaching reasoning to llms without rl or distillation. *Preprint*, arXiv:2507.09850. Accepted at the 2nd AI for Math Workshop at ICML 2025.

Michael N. Forster. 1993. Hegel's dialectical method.

Anamaria Fălăuş. 2020. Negation and alternatives: Interaction with focus constituents. In *The Oxford Handbook of Negation*. Oxford University Press.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *Preprint*, arXiv:2503.01307.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Leon Lukas Hammerla, Andy Lücking, Carolin Reinert, and Alexander Mehler. 2025. D-neg: Syntax-aware graph reasoning for negation detection. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1432–1454, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Laurence R. Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L. Griffiths, Max Kleiman-Weiner, Jiawei Han, Asli Celikyilmaz, and Yulia Tsvetkov. 2025. Cognitive foundations for reasoning and their manifestation in llms. *Preprint*, arXiv:2511.16660.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Manfred Krifka. 2003. The semantics and pragmatics of polarity items.

James Lamb, Nikita Titov, Guolin Ke, wxchan, Laurae, shiyu1994, Qiwei Ye, José Morales, OMOTO Tsukasa, david-cortes, Belinda Trotta, Zhuyi Xue, Oliver Borchert, Ilya Matiach, xuehui, Alberto Ferreira, Nick Miller, Huan Zhang, Chen Yufei, and 11 others. 2025. microsoft/LightGBM. https://github.com/microsoft/LightGBM.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2025. Exploring the reliability of large language models as customized evaluators for diverse NLP tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10325–10344, Abu Dhabi, UAE. Association for Computational Linguistics.

Sara Vera Marjanovic, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stanczak, and Siva Reddy. 2026. Deepseek-r1 thoughtology: Let's think about LLM reasoning. *Transactions on Machine Learning Research*.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jiangbo Pei, Peiyu Liu, Xin Zhao, Aidong Men, and Yang Liu. 2025. Socratic style chain-of-thoughts help LLMs to be a better reasoner. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12384–12395, Vienna, Austria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof Q&A benchmark. In *Proceedings of the First Conference on Language Modeling (COLM)*.

Claudia Schon, Sophie Siebert, and Frieder Stolzenburg. 2021. Negation in cognitive reasoning. In *KI 2021: Advances in Artificial Intelligence*, pages 217–232, Cham. Springer International Publishing.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2025. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1620–1634.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, and Hiroaki Funayama. 2023. Assessing step-by-step reasoning against lexical negation: A case study on syllogism. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 14753–14773, Singapore. Association for Computational Linguistics.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025a. Reasoning models know when they're right: Probing hidden states for self-verification. In *Proceedings of the Second Conference on Language Modeling (COLM)*.

Mike Zhang, Johannes Bjerva, and Russa Biswas. 2025b. Follow the path: Reasoning over knowledge graph paths to improve llm factuality. *Preprint*, arXiv:2505.11140.

Yifan Zhang and Team Math-AI. 2025. American invitational mathematics examination (aime) 2025.

A Human Evaluation

Across tasks, inter-human agreement is highest on gpqa ($\kappa = 0.77$), followed by aime25 ($\kappa = 0.63$), and lowest on nemo ($\kappa = 0.50$). Human-LLM agreement shows a related but not identical pattern. On aime25, agreement with the LLM ranges from $\kappa = 0.36$ to 0.45; on gpqa, it ranges from $\kappa = 0.35$ to 0.46; and on nemo, both annotators yield $\kappa = 0.25$. Thus, although gpqa has the highest inter-human agreement, its human-LLM agreement is not uniformly higher than that of aime25. By contrast, nemo yields the lowest human-LLM agreement for both annotators (Table 2). A similar pattern appears across models. qwen3:32b shows the strongest human-LLM alignment ($\kappa = 0.65$ for both annotators) and also perfect inter-human agreement ($\kappa = 1.00$). deepseek-r1:32b also exhibits relatively strong human-LLM agreement ($\kappa = 0.53$ for both annotators), paired with high human-human agreement ($\kappa = 0.848$). In contrast, gpt-oss:120b and gpt-oss:20b show comparatively weak human-LLM agreement despite moderate to high inter-human agreement on the same samples (Table 3).

Agreement Metric	aime25	gpqa	nemo
H1-LLM (κ)	0.45	0.35	0.25
H2-LLM (κ)	0.36	0.46	0.25
H-H (κ)	0.63	0.77	0.50

Table 2: Agreement (Cohen's κ) across reasoning traces on different benchmarks.

Model	H1-LLM (κ)	H2-LLM (κ)	H-H (κ)
gpt-oss:120b	0.07	0.28	0.88
deepseek-r1:70b	0.41	0.37	0.54
deepseek-r1:32b	0.53	0.53	0.85
magistral:24b	0.24	0.27	0.31
gpt-oss:20b	0.12	0.09	0.69
qwen3:32b	0.65	0.65	1.00

Table 3: Agreement (Cohen's κ) across reasoning traces from different LLMs.

B Preliminary Analysis (Refining vs. Corrective Negation Frequencies)

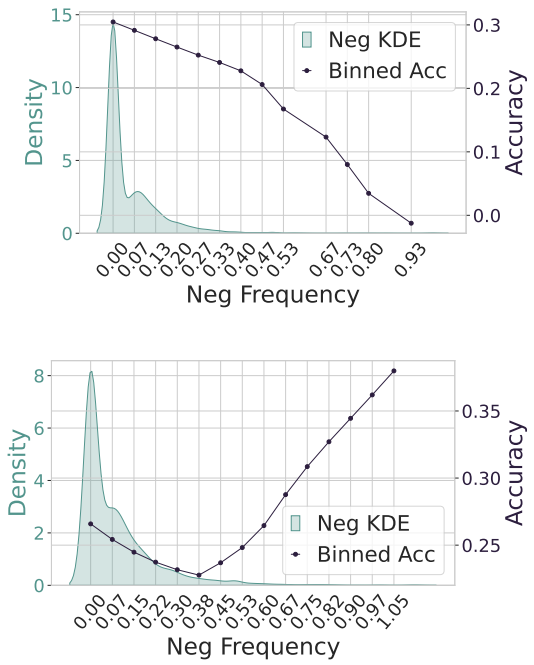
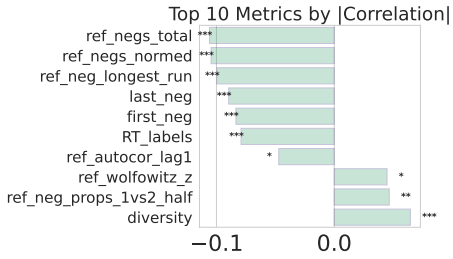


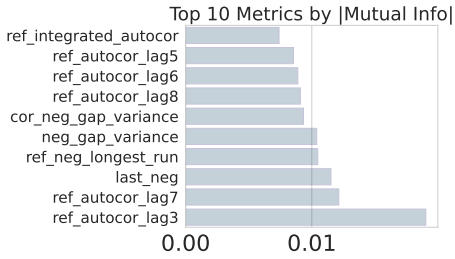
Figure 9: Aggregated results by datasets and models: binned accuracies and kernel density estimates of negation frequencies (negations per reasoning step), smoothed using LOWESS. Corrective negation is shown on the top, and refining negation on the bottom.

To investigate the contrasting behavior of refining and corrective negations, we bin all reasoning traces across datasets according to the frequency of each negation type and compute the mean model accuracy within each bin. We then apply LOWESS smoothing to the resulting trends to obtain robust trajectories. The patterns reveal a clear divergence between the two types: accuracy initially declines for both, but the drop is substantially steeper for corrective negations. At higher frequencies, refining negations exhibit a recovery and eventually show a positive association with model accuracy, whereas corrective negations continue to decline monotonically. This indicates that increased use of refining negation correlates with improved model performance, while corrective negation does not. These effects are further supported by single-metric analyses reported in Appendix C. The full smoothed curves are shown in Figure 9.

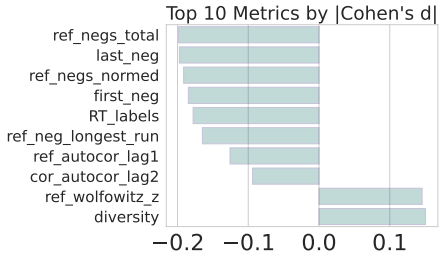
C Single-metrics (negative) results



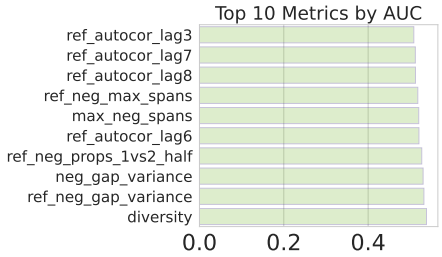
(a) Pairwise Spearman correlations between correctness (binary) and the proposed metrics. Significance levels are marked with asterisks: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.



(b) Mutual information between correctness (binary) and the proposed metrics, with continuous metrics discretized by binning.



(c) Distributional differences of the proposed metrics between correct and incorrect samples, quantified by Cohen's d.



(d) Discriminative ability of metrics between correct and incorrect samples, measured by AUC.

Figure 10: Comparison of different statistical measures assessing the relationship between the proposed metrics and binary correctness across all datasets. Each subplot reports the top 10 metrics ranked by the respective measure.

D Multi-metrics Result

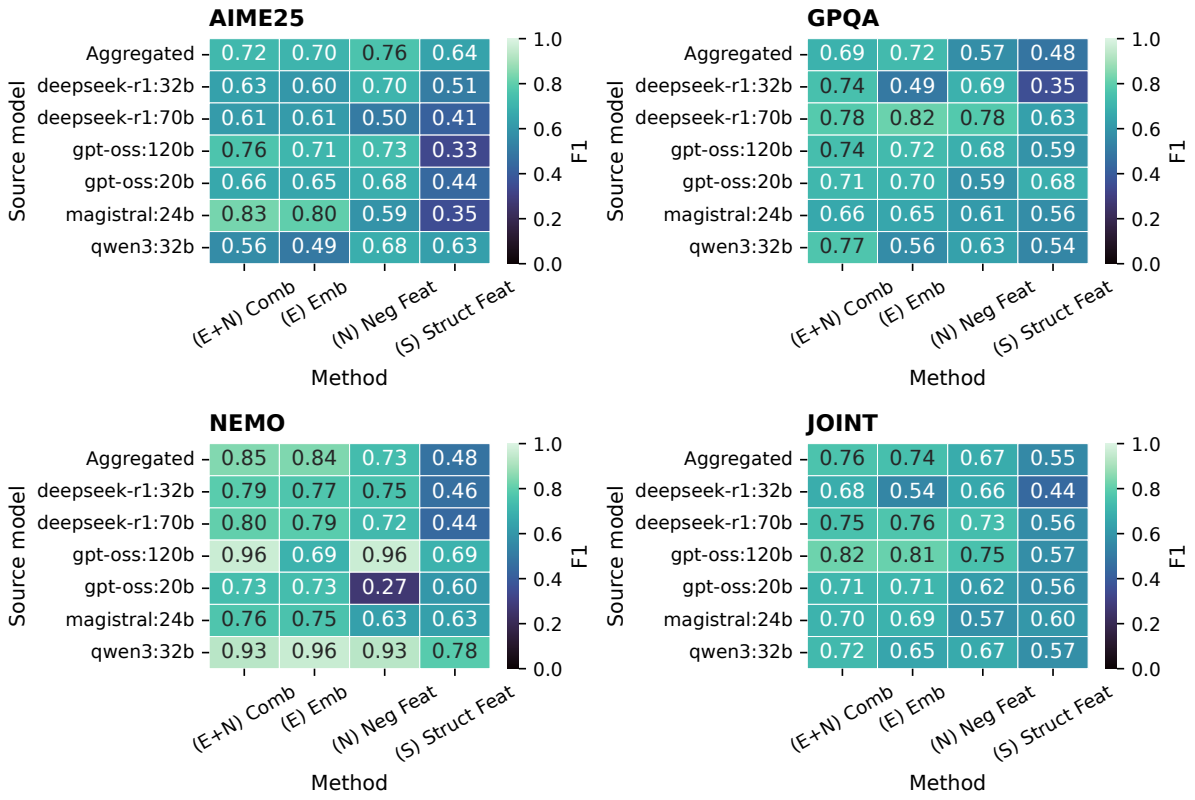


Figure 11: Five-fold CV binary F1 scores for correctness prediction on each dataset and the merged dataset, grouped by source model and feature setting.

E SHAP Analysis

E.1 Predicting Model Correctness (SHAP)

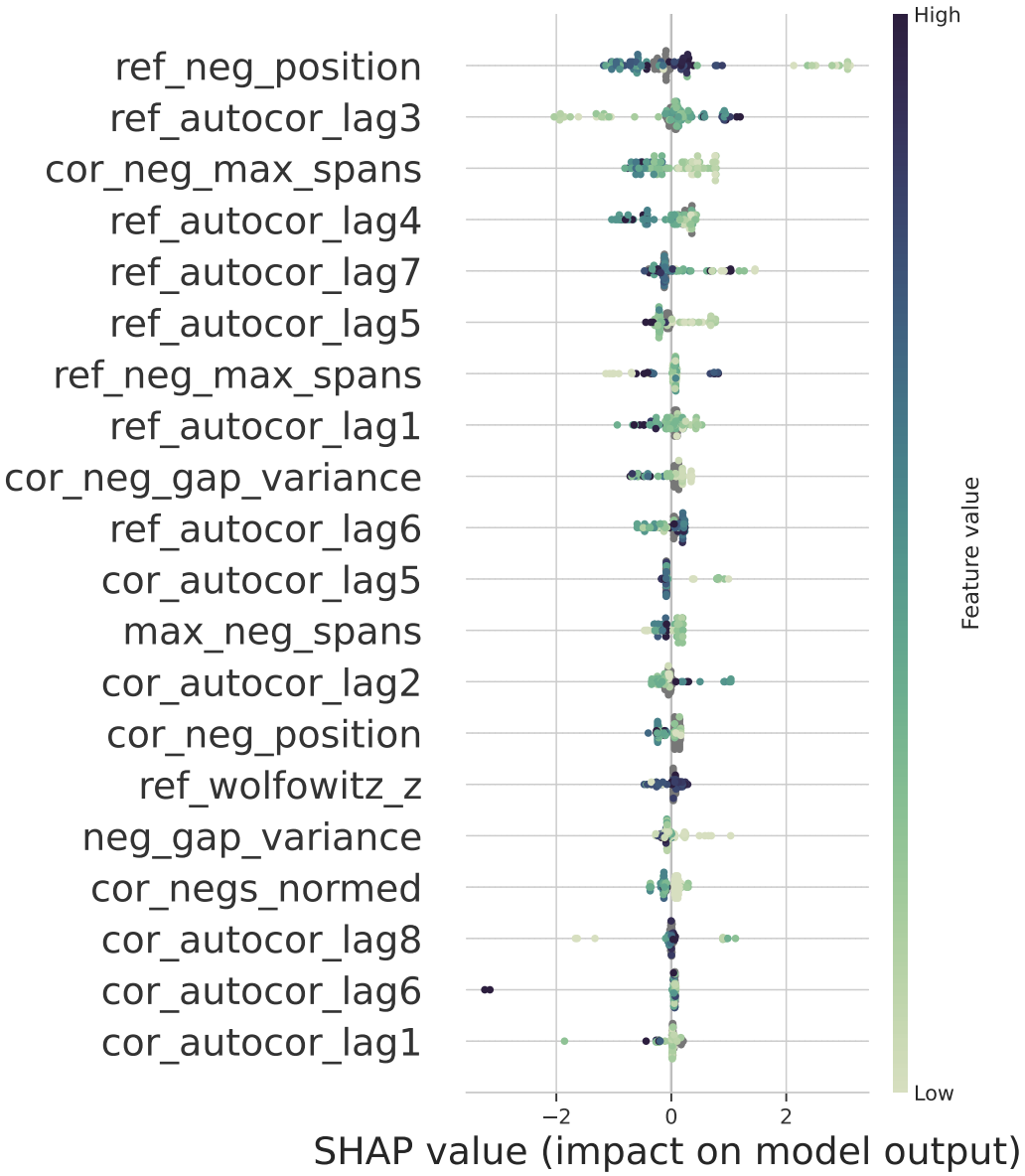


Figure 12: SHAP values showing the impact of negation metrics on predicting response correctness from reasoning traces using gradient boosting. Predictions were made using aggregated data from all three datasets.

E.2 Predicting Identity (LLM)

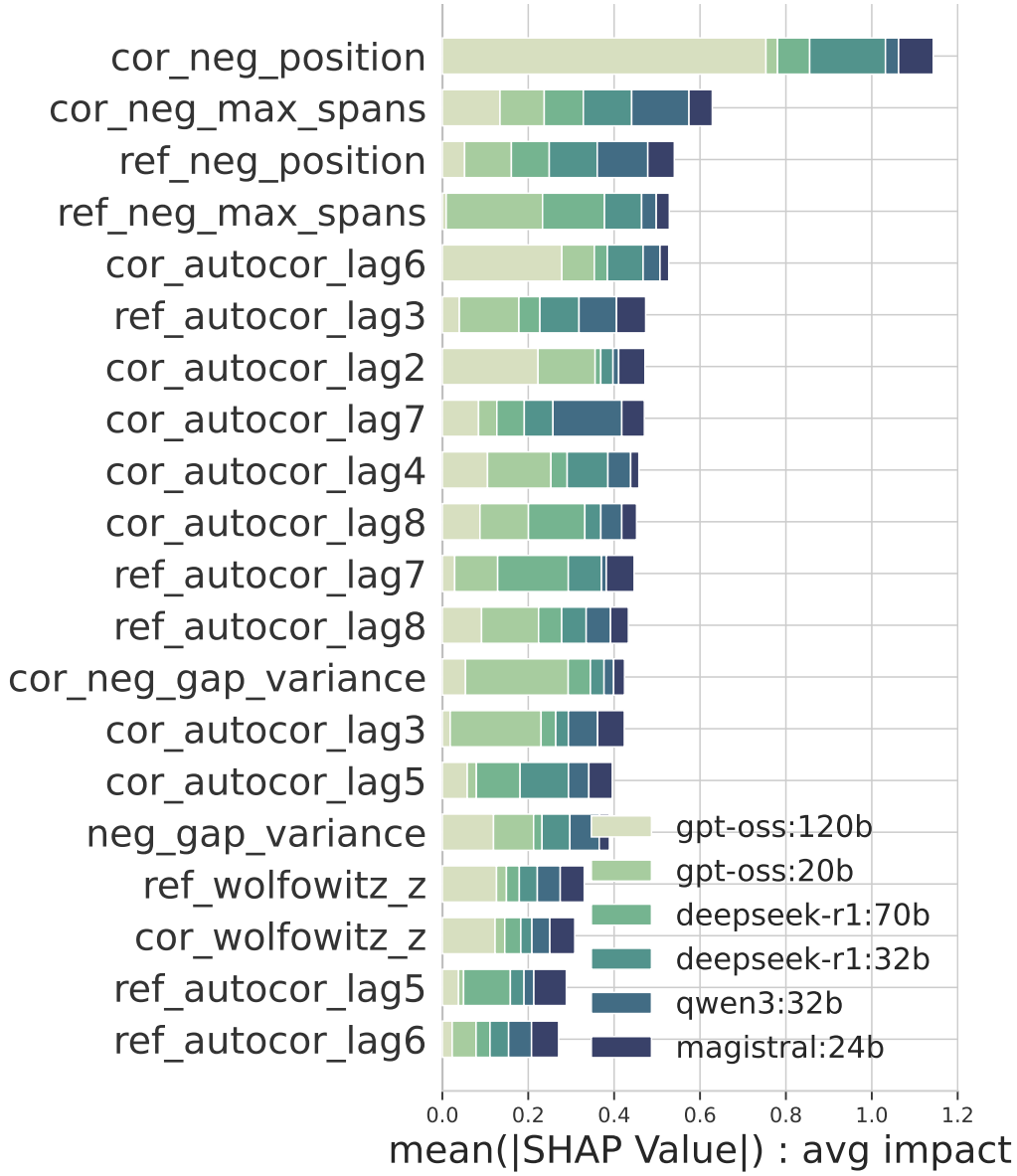


Figure 13: Mean SHAP values showing the contribution of each feature to the model output when predicting which model produced a reasoning trace, based on its negation quantification (top 20 features displayed).

E.3 Predicting Identity (Human-LLM)

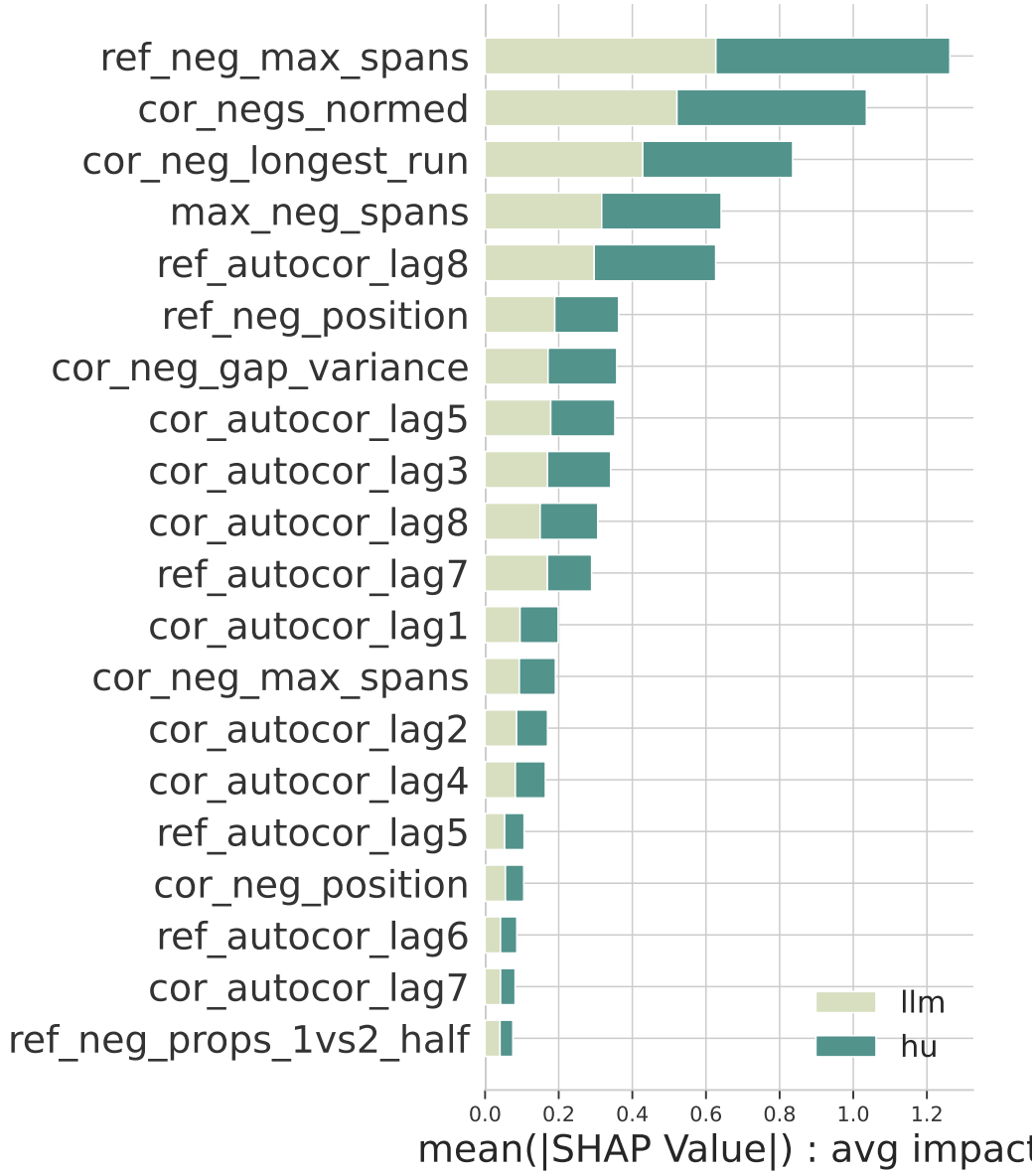


Figure 14: Mean SHAP values showing the contribution of each negation feature to the model output when classifying human and LLM reasoning traces, based on their negation quantification (top 20 features displayed).

F Prompts

F.1 Dataset - Model Benchmark (Step 1)

```
You are given a multiple-choice question with four options. Only one option is correct.

1. Provide a brief explanation for your choice.
2. Additionally, add your answer in JSON format:
   {"answer":ANSWER_ID}
   where ANSWER_ID is the number (0, 1, 2, or 3) corresponding to the correct option.

---

Question:
[question]

---

Options:
0: [option1]
1: [option2]
2: [option3]
3: [option4]
```

Figure 15: Prompt for creating reasoning traces and model responses for the gpqa benchmark.

```
You are tasked with solving a math question where the answer is guaranteed to be an integer.

1. Provide a concise explanation of your solution.
2. Verify that your answer is an integer, as required by the question.
3. Additionally, add your answer in JSON format: {"answer": RESULT}, where RESULT is an integer solution to the provided
↪  question.

---

Question:
[question]
```

Figure 16: Prompt for creating reasoning traces and model responses for the aime25 and nemo benchmark.

F.2 Dataset - Human Error (Step 2)

```
You are given a **correct human reasoning trace**.
Your task is to generate a **perturbed version of that reasoning trace** that:
* stays *close* to the original (similar structure, similar steps, similar flow)
* includes **human-like reasoning slips** (minor misreads, small leaps, overlooked details, mis-prioritizations)
* modifies the trace only slightly --- **do not rewrite it from scratch**
* introduces perturbations according to a severity scale:
  * **1 = very light distortions** (tiny misinterpretations, small skipped details)
  * **2 = moderate distortions** (noticeable but still realistic human mistakes)
  * **3 = strong distortions** (multiple compounded human-like errors, while still keeping the trace coherent)
**Important:**
* The output should **only be the perturbed reasoning trace itself**, do not include any kind of observation or explanation.
* **Do not** produce a final answer or solution --- the task is only to perturb the reasoning trace, not to solve the problem.
* The perturbed trace must remain recognizably similar to the original.
* The perturbation must be based *directly* on the original trace, not generated fresh.
```

Figure 17: Prompt for introducing human reasoning errors.

F.3 Dataset - Annotation (Step 3)

````
**Instruction:**
You are an annotator for reasoning chains generated by language models. Your task is to:
---
#### **1. Chunk the reasoning text**
* Break the reasoning text into **discrete reasoning steps**.
* Each step can contain **one or more sentences** representing a coherent thought, inference, or claim.
* Number the steps sequentially starting from 1.
---
#### **2. Detect corrective or refining negations**
Annotate only negations that:
* **CORRECTIVE_NEGATION**: Directly contradict or correct one or more earlier steps.
* **REFINING_NEGATION**: Narrow, qualify, or improve prior reasoning without fully contradicting it.
**Important:**
* Negations may appear **anywhere within a step**, even if the step contains multiple sentences.
* Only annotate negations that **modify previous reasoning to improve accuracy, precision, or clarity**.
* Ignore general negations that are purely factual, descriptive, or do not impact prior reasoning.
---
#### **3. Annotation requirements**
For each detected negation, provide:
* `type`: `CORRECTIVE_NEGATION` or `REFINING_NEGATION`
* `step`: Step number where the negation occurs
* `text`: Exact negated phrase in that step
* `refines_steps` or `corrects_steps`: List of prior step numbers affected
* If a single step contains **multiple corrective/refining negations**, list each separately.
* A negation can refer to **one or more previous steps**, and this should be reflected in the `refines_steps` or `corrects_steps` lists.
---
### **Output format (JSON-like)**
```json
{
  "steps": [
    "Step 1 text (can be multiple sentences)",
    "Step 2 text",
    ...
  ],
  "negations": [
    {
      "type": "CORRECTIVE_NEGATION",
      "step": 2,
      "text": "",
      "corrects_steps": [1]
    },
    {
      "type": "REFINING_NEGATION",
      "step": 3,
      "text": "",
      "refines_steps": [1]
    }
  ]
}
```
---
### **Example 1: Object Color (Extended)**
**Reasoning chain:**
*"The object appears red from one angle. Under direct light, it seems vibrant and uniform. But this may not be accurate because shadows and reflections can distort perception. Furthermore, it is not necessarily completely red; some areas show orange hints. Considering these observations, the object might actually be a mix of red and orange. Therefore, it is safer to describe the object as mostly orange with red undertones."*
**Annotation:**
```json
{
  "steps": [
    "The object appears red from one angle. Under direct light, it seems vibrant and uniform.",
    "But this may not be accurate because shadows and reflections can distort perception.",
    "Furthermore, it is not necessarily completely red; some areas show orange hints.",
    "Considering these observations, the object might actually be a mix of red and orange.",
    "Therefore, it is safer to describe the object as mostly orange with red undertones."
  ],
  "negations": [
    {
      "type": "CORRECTIVE_NEGATION",
      "step": 2,
      "text": "this may not be accurate",
      "corrects_steps": [1]
    },
    {
      "type": "REFINING_NEGATION",
      "step": 3,
      "text": "not necessarily completely red",
      "refines_steps": [1]
    }
  ]
}
```
...
### **Example n: xxx (Extended)**
...
````

Figure 18: Prompt for reasoning trace segmentation and discourse-level negation type annotation.