Lexical Hints of Accuracy in LLM Reasoning Chains

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning Large Language Models (LLMs) with reinforcement learning to produce an explicit Chain-of-Thought (CoT) before answering produces models that consistently raise overall performance on code, math, and general-knowledge benchmarks. However, on benchmarks where LLMs currently achieve low accuracy, such as Humanity's Last Exam (HLE), they often report high self-confidence, reflecting poor calibration. Here, we test whether measurable properties of the CoT provide reliable signals of an LLM's internal confidence in its answers. We analyze three feature classes: (i) CoT length, (ii) intra-CoT sentiment volatility, and (iii) lexicographic hints, including hedging words. Using DeepSeek-R1 and Claude 3.7 Sonnet on both Humanity's Last Exam (HLE), a frontier benchmark with very low accuracy, and Omni-MATH, a saturated benchmark of moderate difficulty, we find that lexical markers of uncertainty (e.g., *guess*, *stuck*, *hard*) in the CoT are the strongest indicators of an incorrect response, while shifts in the CoT sentiment provide a weaker but complementary signal. CoT length is informative only on Omni-MATH, where accuracy is already high ($\approx 70\%$), and carries no signal on the harder HLE ($\approx 9\%$), indicating that CoT length predicts correctness only in the intermediate-difficulty benchmarks, i.e., inside the model's demonstrated capability, but still below saturation. Finally, we find that uncertainty indicators in the CoT are consistently more salient than high-confidence markers, making errors easier to predict than correct responses. Our findings support a lightweight post-hoc calibration signal that complements unreliable self-reported probabilities and supports safer deployment of LLMs.

1 Introduction

Large Language Models (LLMs) can be fine-tuned with reinforcement learning to produce a Chain-of-Thought (CoT) before delivering their final response to improve their general performance (Jaech et al., 2024; Korbak et al., 2025; Muennighoff et al., 2025; Wei et al., 2022). The fine-tuning rewards are based both on the final correctness of the response—to improve capabilities—and on adherence to the CoT format, which may, for example, include safety-related specifications to enhance readability and alignment (Guan et al., 2024; Guo et al., 2025). Although LLMs have shown a strong performance increase on demanding benchmarks, such as GPQA (Rein et al., 2023), SWE-bench (Jimenez et al., 2024), and FrontierMath (Glazer et al., 2024), they often report very high confidence while still obtaining relatively low overall accuracy, reflecting poor calibration (Marjanović et al., 2025; Phan et al., 2025; Wei et al., 2024). This miscalibration masks silent failure modes and undermines the reliability of LLMs in open-ended settings (Chen et al., 2025a; Ke et al., 2025).

In addition to the poorly calibrated self-reported confidence (Phan et al., 2025), the confidence of LLMs in their final responses can be assessed via sampling agreement techniques that are computationally expensive (Lyu et al., 2025; Cherian et al., 2024; Kadavath et al., 2022), or by monitoring internal representations, which requires access to the model weights (Baek & Tegmark, 2025; Templeton et al., 2024; Zou et al., 2023). In contrast, the CoT of some LLMs is readily available with its final response, offering a potential proxy for assessing the model's internal confidence (Baker et al., 2025; Lindsey et al., 2025). Although it has been shown that an LLM's CoT may not always faithfully reflect the model's reasoning process and may be difficult to interpret (Chen et al., 2025b; Kambhampati et al., 2025; Korbak et al., 2025; Stechly et al., 2025), there is

some evidence that specific CoT properties predict the accuracy of the final response (Jiang et al., 2025; Wu et al., 2025). For example, there is mixed evidence on how informative the length of the CoT is for predicting response accuracy (Ballon et al., 2025; Chen et al., 2024; Shojaee et al., 2025; Su et al., 2025; Wang et al., 2025).

In this paper, we investigate whether Chain-of-Thought (CoT) provides a reliable proxy for an LLM's internal confidence by systematically analyzing three classes of features: (i) CoT length, aligning with prior work, (ii) intra-CoT sentiment volatility, and (iii) lexicographic indicators, such as hedging words. Whereas most existing studies focus narrowly on length-based correlations, our contribution is to extend this analysis with sentiment and lexical markers. We evaluate these features on two contrasting benchmarks: Omni-MATH, where state-of-the-art models already perform well, and Humanity's Last Exam (HLE), which remains highly challenging. The results show that CoT length as well as sentiment dynamics has some predictive power in the saturated Omni-MATH setting but provides little signal on HLE, while lexical cues consistently yield informative calibration signals across both tasks.

2 Methods

2.1 Datasets and Evaluation

To analyze the lexicographic and linguistic characteristics of LLM reasoning, we collect the CoT and final responses from DeepSeek-R1 (Guo et al., 2025) and Claude Sonnet 3.7 (Anthropic, 2025a) on two large-scale benchmarks of varying difficulty: Omni-MATH (Gao et al., 2024) and HLE (Phan et al., 2025). We choose DeepSeek-R1 and Claude 3.7 Sonnet because their complete CoTs are fully available, and the serial test-time compute protocol that produces them—including the configurable maximum-token cap—is publicly documented (Guo et al., 2025; Anthropic, 2025b). Although the two models achieve comparable benchmark scores, Claude 3.7 Sonnet employs more extensive alignment techniques, such as constitutional fine-tuning (Bai et al., 2022), which may affect its calibration and the uncertainty signals visible in its CoT (Zhu et al., 2023).

Starting from the original $3,000$-question HLE dataset (Phan et al., 2025), we first remove 316 multimodal items. From the remaining $2,684$ questions, we then exclude 412 multiple-choice items, resulting in $2,088$ open-ended questions used in our analysis. The official benchmark results report that DeepSeek-R1 and Sonnet 3.7 achieve 8.6% (calibration error: 81.4%) and 8.9% (calibration error: 88.3%) accuracy on the full (text-only for DeepSeek-R1) version of the benchmark, respectively (for AI Safety, 2025). Omni-MATH comprises $4,428$ mathematics problems ranging from basic arithmetic to advanced algebra and geometry. On this benchmark, reasoning models, such as OpenAI o3-mini, achieve over 70% accuracy and even surpass 80% accuracy for algebra and calculus in a high reasoning effort setting (Ballon et al., 2025).

Omni-MATH and HLE are intentionally contrasting benchmarks: Omni-MATH offers $4,428$ tiered math problems with automatic, high-fidelity grading (Gao et al., 2024), allowing us to observe CoT behavior across a continuous, well-calibrated difficulty spectrum. HLE, by contrast, poses $2,088$ multidisciplinary, open-ended questions that push models to the edge of their reasoning abilities, often revealing brittle generalization and poor calibration under genuine uncertainty (Phan et al., 2025).

For both data sets, we prompt each model with a single, standardized template designed to elicit a clear explanation, a concise final answer, and a self-reported confidence score (0%-100%), all the while we log its CoT:

Your response should be in the following format:

Explanation: your explanation for your final answer

Exact Answer: your succinct, final answer

Confidence: your confidence score between 0 % and 100 % for your answer

Solve the following problem:

Additionally, for Claude Sonnet 3.7, we set its max tokens to 128k and its thinking budget to 100k tokens. DeepSeek-R1 has a max tokens parameter set to 128k, which includes the CoT tokens.

To grade the final responses, we employ a customized grading strategy that accommodates the unique challenges and features of each benchmark. Omni-MATH's responses were automatically scored using Omni-Judge (Gao et al., 2024), an automatic mathematical grader, whose alignment with human judgment has been explicitly validated for Omni-MATH (Ballon et al., 2025; Gao et al., 2024; Verga et al., 2024).

We prompted Omni-Judge as follows:

OBJECTIVE

You are tasked with evaluating the correctness of a student's answer. Below, you are provided with a problem, a reference answer, and a student's answer. You should assess whether the student's answer captures the same meaning as the reference answer, even when expressed with different wording or format.

Your tasks include:
A. Identify Mathematical or Notational Equivalence.
B. Conclude with a brief explanation as to why the student's output is correct or incorrect.

 # RESPONSE: MARKDOWN REPORT #
Student Final Answer
Extract the student's final answer, which is enclosed in "\boxed{}".

Equivalence Judgement
Whether the student's answer shares the same meaning with the reference answer. (TRUE or FALSE)

Justification
Conclude with a brief explanation as to why the student's answer is correct or incorrect.

ATTENTION
- The reference answer is ALWAYS correct. You should carefully judge whether the student gives the same answer as the reference answer.
- The answer is FALSE even if the student's final answer is almost correct with a minor mistake.
- The answer is contained within the "boxed" section, so you can focus solely on comparing the content in the student's answer box with the reference answer, without needing to consider the intermediate steps.
- Add "=== report over ===" at the end of the report.

For HLE, each answer is independently assessed by both a human expert and a different LLM. We use OpenAI o3-mini to avoid preference leakage, which can occur when the same model is used for both answer generation and evaluation, potentially resulting in more lenient or biased assessments of its own responses (Li et al., 2025). The agreement between human and automated grades is high: 93.9% for Claude 3.7 Sonnet

and 93.4% for DeepSeek-R1, corresponding to Cohen's kappa values of 0.88 and 0.87, indicating near-perfect alignment. For further analysis, we include only HLE items where human and OpenAI o3-mini evaluations agree. Finally, to assess model calibration, we group the outputs into decile confidence intervals (0% - 10%, 10% - 20%, etc.) and calculate the mean absolute difference between the average reported confidence of each bin and its empirical precision.

Both DeepSeek-R1 and Claude Sonnet 3.7 perform similarly across the two benchmarks and in line with their official benchmark results. Note that a small derivation is to be expected as Bowyer et al. (2025) (Bowyer et al., 2025) argue, point estimates on small, specialized tests should be reported with confidence intervals to avoid underestimating uncertainty (e.g., error bars or significance tests). On HLE, Claude Sonnet 3.7 achieves a slightly higher accuracy than DeepSeek-R1 (9.2% vs. 8.9%), whereas on Omni-MATH, DeepSeek-R1 outperforms Claude Sonnet 3.7 (72.5% vs. 69.1%). In both cases, Claude Sonnet 3.7 displays the highest calibration error. The large difference in accuracy between Omni-MATH and HLE highlights the high complexity of HLE. A global overview of accuracies and calibration errors is presented in Table 1. A full breakdown, categorized by difficulty level (only for Omni-MATH), is provided in Appendix A.

Table 1: Overall accuracy and calibration error for DeepSeek-R1 and Claude 3.7 Sonnet on the HLE and Omni-MATH benchmarks. While both models perform similarly, Claude Sonnet 3.7 is slightly more accurate on HLE, and DeepSeek-R1 outperforms Claude on Omni-MATH. Claude consistently exhibits higher calibration error across both benchmarks.

Dataset	**Model**	**Accuracy (%)**	**Calibration Error (%)**
HLE	DeepSeek-R1	8.6%	78.1%
	Claude 3.7 Sonnet	9.2%	84.6%
Omni-MATH	DeepSeek-R1	72.5%	20.1%
	Claude 3.7 Sonnet	69.1%	29.0%

2.2 Intra-CoT sentiment Volatility in the Chain-of-Thought

The ability to regulate emotions, such as maintaining composure or overcoming moments of doubt, has been linked to improved cognitive outcomes in humans (Schutz et al., 2004; Schutz & Davis, 2000; Schutz et al., 2008; Harley et al., 2021). Motivated by this, we examine the impact of the Intra-CoT sentiment volatility within the CoT responses on the response correctness. Specifically, we asked whether the model's reasoning began with a confident, optimistic tone and ended in doubt, or if it started hesitantly and grew more optimistic as the answer developed.

To quantify this, we process each CoT through OpenAI o3-mini as an independent sentiment evaluator (Tori et al., 2024). For each rationale, we extract two sentiment scores—the most negative and the most positive sentiment, ranging from -1 (strongly negative) to +1 (strongly positive). We also record the order in which these sentiments appear in the CoT, i.e., whether the response moved from negative to positive (an "upward" direction, scored as 1), from positive to negative ("downward," scored as –1), or remained neutral throughout (scored as 0). We use the following prompt to evaluate each CoT:

> Judge how emotional this text is by reporting on its most positive emotion and its most negative emotion on a scale from -1 (negative) to 1 (positive), and the order in which they appear, i.e. direction in which they appear, if the most negative emotion appears before the most positive emotion the direction is upward so, 1, if inverse is true the direction is -1 if the text is completely neutral and no gradient exists present between the two emotional extremes the direction is 0, picking up on small effects. Always end your response with:
>
> Most positive emotion Score: ...
> Most negative emotion Score: ...
> Direction: ...

Using the sentiment scores provided by OpenAI o3-mini, we define Intra-CoT sentiment volatility as the signed difference between the most positive and most negative sentiment within each CoT. Large differences (Δ) may indicate moments of uncertainty, backtracking, or sudden shifts in confidence, while smaller deltas suggest a more consistent and steady tone throughout the reasoning process.

2.3 Lexicographic Analysis of the Chain-of-Thought

To explore the connection between the linguistic characteristics of model-generated rationales and their accuracy, we perform a lexicographic analysis of CoT responses from all evaluated datasets and models. For each unique word found within a CoT, we calculate its relative accuracy by comparing the success rate of responses containing that word to the average accuracy within the respective dataset. To ensure reliability, we exclude low-frequency words by requiring that each word appear at least a minimum number of times (300 occurrences) in every dataset to be included in the analysis.

2.3.1 Hedging Words and Accuracy

In addition to the general lexicographic analysis, we focus on a specific subset of the lexicon: hedging words. In human communication, hedging—through modal verbs (e.g., might, could), uncertainty adverbs (e.g., possibly, perhaps), and tentative phrases (e.g., it seems that)—serves as a cue indicating lower confidence (Hyland, 1998; Demir, 2018; Lakoff, 1973). We hypothesize that the same might be true for the CoT of LLMs.

Drawing on established taxonomies of uncertainty markers, (Hyland, 1998; Demir, 2018; Lakoff, 1973), we compiled a lexicon of such expressions and computed a hedging rate per, i.e., proportion of sentences in a rationale containing at least one hedging expression CoT. The full hedging lexicon used was:

- Modal and uncertainty verbs: might, may, could, should, would, seems, suggests, appears
- Uncertainty adverbs: possibly, perhaps, likely, unlikely, probably, generally, usually, sometimes, often, tends, somewhat, rather, quite, almost, nearly, virtually, presumably, arguably, relatively, fairly, reasonably, mostly, partially, mainly, primarily, essentially, basically
- Common hedging phrases: it seems that, it appears that, it suggests that, it is possible that, it is likely that
- Additional qualifiers: in part, to some extent

3 Results

3.1 Relationship Between CoT Length and Accuracy

To assess how the length of the CoT relates to response correctness, we analyze accuracy as a function of the number of words in each CoT rationale. We observe a downward trend in the Omni-MATH benchmark: accuracy declines as CoT length increases (see Figure 1). Claude 3.7 Sonnet proves more robust to longer reasoning chains than DeepSeek-R1, losing on average 3% of accuracy per additional 1,000 CoT words,

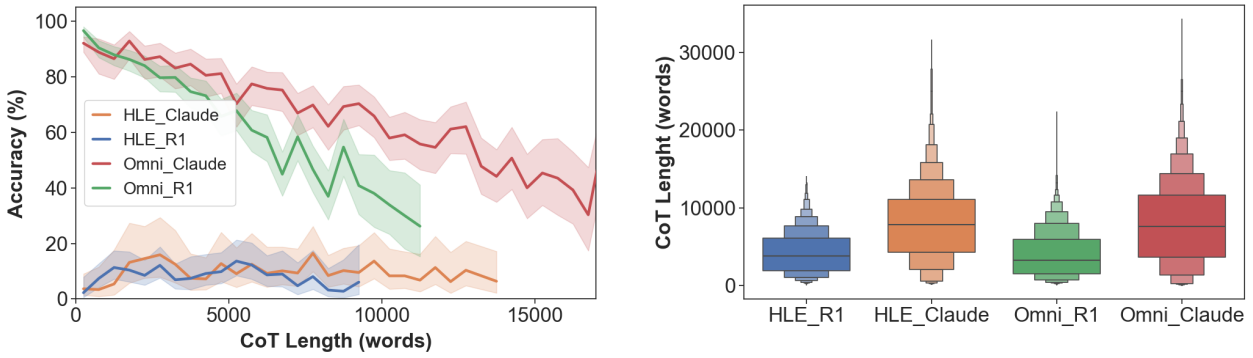


Figure 1: Accuracy as a function of CoT length (left) and the distribution of CoT lengths (right) for each model and benchmark. On Omni-MATH, accuracy declines with longer responses, especially for DeepSeek-R1; in HLE, accuracy remains low and variable regardless of length. CoT length bins with fewer than 30 samples were excluded from the accuracy analysis on the left. Claude Sonnet 3.7 generally produces longer chains-of-thought than DeepSeek-R1, likely a reflection of different reasoning allowances.

compared to 6.2% for DeepSeek-R1. These results echo the findings of (Ballon et al., 2025), who report that o3-mini-high loses about 0.8 % accuracy per 1,000 additional reasoning tokens on Omni-MATH benchmark. By contrast, in HLE, we find that the CoT length shows no discernible pattern, accuracy stays low, and fluctuates widely regardless of CoT length. This suggests that CoT length predicts correctness only for benchmarks of intermediate difficulty—that is, tasks within the model's demonstrated capabilities but not yet saturated. These findings are consistent with previous work showing a negative relationship between CoT length and response accuracy on Omni-MATH (Ballon et al., 2025), as well as the disappearance of this effect on the most challenging benchmarks (Shojaee et al., 2025; Opus & Lawsen, 2025). In general, we find that Claude Sonnet 3.7 generally produces longer CoTs than DeepSeek-R1, which may be due to architectural choices or differences in reinforcement learning post-training (Muennighoff et al., 2025; Guo et al., 2025).

3.2 Intra-CoT sentiment Volatility and Reasoning Dynamics

Figure 2 shows the relationship between Intra-CoT sentiment volatility within CoT and response correctness. We observe that a parabolic pattern emerges for Omni-MATH, with accuracy peaking at a volatility value of approximately 0.1. This suggests that CoTs with a consistent or slightly uplifting sentiment are associated with the highest accuracy. In contrast, no such pattern is observed for HLE, where accuracy remains uniformly low regardless of the degree of sentimental variation. To rule out the possibility that a nonzero mean (baseline) emotion score was driving this effect, we repeated the volatility-accuracy analysis on mean-centered sentiment values. After centering, we found the same parabolic trend for Omni-MATH and the same flat pattern for HLE. While both reasoning models display a predominantly neutral CoT, they exhibit markedly different sentimental footprints. Claude 3.7 Sonnet's responses tend to be slightly optimistic in tone, becoming more positive over the course of their reasoning, while DeepSeek's rationales often exhibit a slight negative shift. This is marked by the more positive Intra-CoT sentiment volatility of Claude 3.7 Sonnet and DeepSeek-R1, which has more weight on the negative side in Figure 2(right).

3.3 Lexical Hints of Uncertainty

To scrutinize how specific words reflect or undermine a model's underlying confidence in response correctness, we construct a lemmatized lexicon from the CoT. We exclude rare or highly exclusive words from the lemmatization process, retaining only those appearing in at least 300 CoT responses for each model and benchmark. We distinguish between lemmatized words that consistently harm or boost accuracy. For ease of reading, we use words to refer to the lemmas in the remainder of the paper.

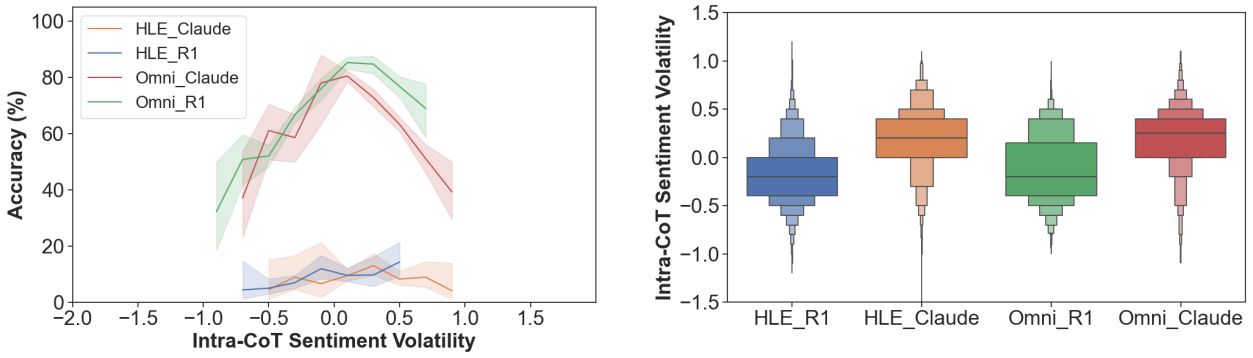


Figure 2: Accuracy as a function of Intra-CoT sentiment volatility (left) and its distribution (right) for each model and benchmark. On Omni-MATH, accuracy peaks for CoTs with low or slightly positive volatility; for HLE, accuracy is uniformly low. Claude 3.7 Sonnet generally produces more positive-toned reasoning, while DeepSeek-R1 leans more to a slightly negative flow of emotion. Bins with fewer than 30 samples were excluded from the left figure.

Interestingly, many of the words most strongly linked to reduced accuracy, such as *complexity*, *guess*, *stuck*, *hard*, *likely*, *probably*, *possibly*, and *complex*, are intuitively associated with task difficulty and cognitive overload, mirroring language that humans use to express uncertainty or task difficulty (Renkl, 1997; Hyland, 1998; Demir, 2018; Lakoff, 1973). This suggests a degree of linguistic convergence in the signaling of uncertainty between reasoning models and humans. As shown in Figure 3, the top 30 words most detrimental to accuracy also include terms conveying confusion or lack of direction, such as *information*, *miss*, *depend*, *beyond*, *help*, and *direction*. A full overview of these lexical hints of uncertainty, i.e., "harmful" words can be found in Appendix E.

In contrast, only the word "so" appears to have consistently boosted accuracy across all datasets (see Figure B). However, when considering its confidence interval, this apparent effect is likely due to random variation rather than a true underlying pattern. As a general rule, the words with the highest mean relative accuracy tend to be those characteristic of mathematical reasoning, including *equivalent*, *solve*, *therefore*, *substitute*, *equal*, *equation*, *calculation*, *formula*, *compute*, and *coefficient*. This is an artefact of both Claude and DeepSeek demonstrating above-average performance on mathematical questions within HLE, achieving accuracies of 10% and 9.1%, respectively.

3.3.1 Hedging

We continue by looking at a subset of our full lexicon, hedging words. Hedging words are typically used by humans to mark doubt or a lack of knowledge (Hyland, 1998; Demir, 2018; Lakoff, 1973); the same could be true for LLM-generated CoT.

Across the four model benchmark pairs, we observe a negative relationship between hedging rate and accuracy: as the proportion of hedged sentences increases, the accuracy tends to decline (Figure 4). This trend is most pronounced for Omni-MATH, where the correlations are $r = -0.24$ ($p < 0.001$) for DeepSeek and $r = -0.14$ ($p < 0.001$) for Claude 3.7 Sonnet, indicating small but statistically detectable negative associations under the null hypothesis of zero correlation. For HLE, the correlations are weaker: $r = -0.10$ ($p < 0.001$) for DeepSeek-R1, indicating a statistically significant deviation from zero, and $r = -0.04$ ($p = 0.06$) for Claude 3.7 Sonnet, which does not differ significantly from zero at conventional thresholds. Overall, while the negative association is consistent in direction, the effect sizes are small.

Moreover, we observe that hedging is more prevalent when models address the more challenging HLE benchmark. In Omni-MATH, the median hedging rate is about 10%, which means that roughly one in ten sentences contains a hedging expression. For HLE, this median rises to about 20%. This pattern, shown in

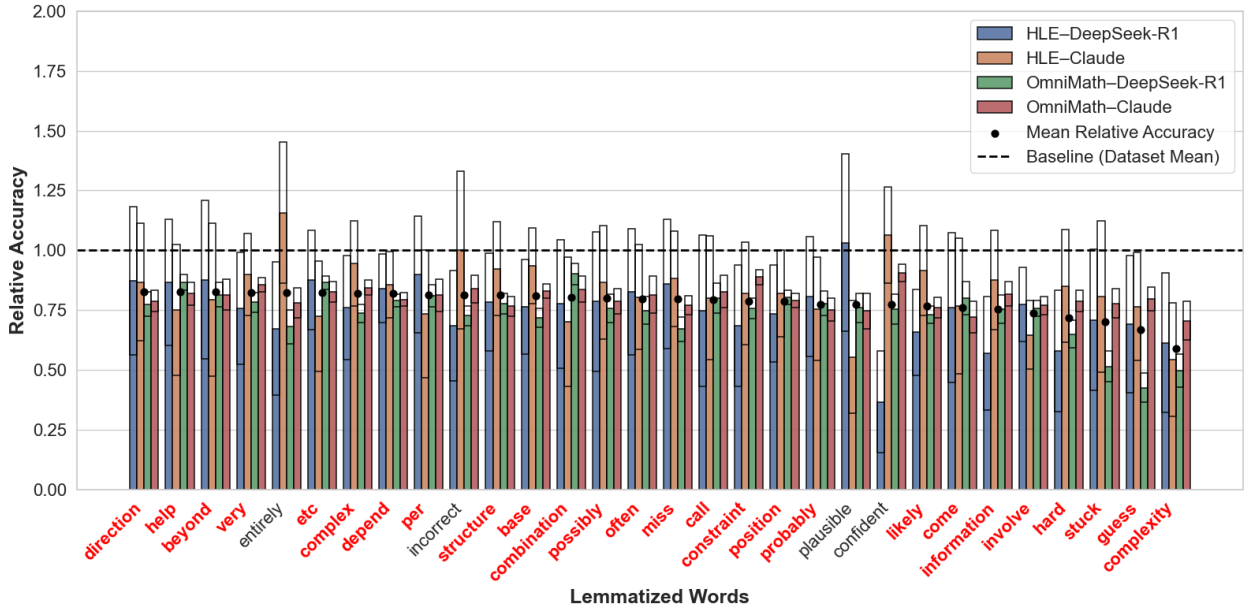


Figure 3: The relative accuracy per dataset/model is shown as a barplot on which words that consistently reduce accuracy have been highlighted in red. Moreover, the dataset mean is marked by a dashed line, and the mean relative accuracies by black dots. Remarkably, many of the words most strongly associated with reduced accuracy are typically linked to uncertainty, cognitive overload, or confusion (e.g., complexity, guess, stuck, hard, likely, possibly, depend, help, direction). These patterns mirror human language for expressing task difficulty and suggest linguistic convergence between models and humans.

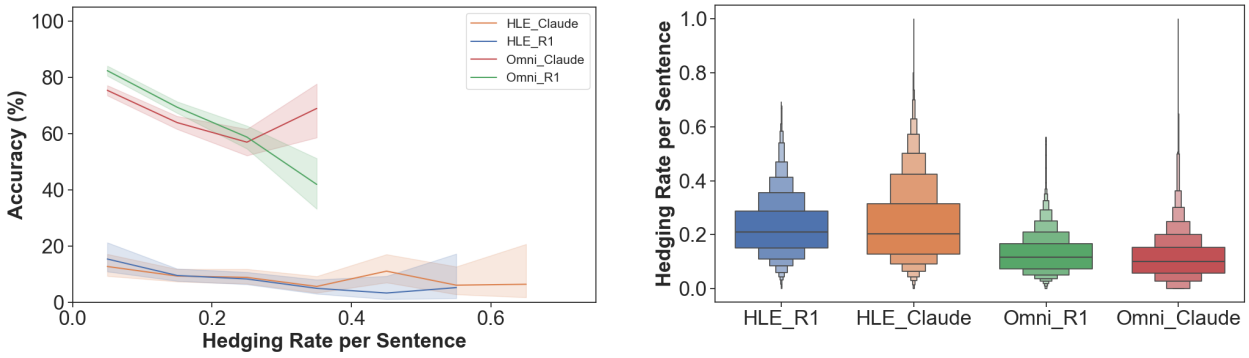


Figure 4: Accuracy as a function of hedging rate (left) and the distribution of hedging rates (right) for each model and benchmark. Higher hedging rates are associated with lower accuracy, especially on Omni-MATH, though the effect is small. Hedging is more common in HLE, with a median rate of 20%, compared to 10% for Omni-MATH, reflecting increased linguistic uncertainty on more challenging tasks. As before, bins with fewer than 30 samples were excluded from the left figure.

Figure 4, suggests that as task difficulty increases, models produce more linguistic signals of uncertainty in their reasoning.

3.4 Prediction Accuracy

The question now remains whether these lexicographic features can be used to predict/anticipate the correctness of a reasoning model's final response. To assess which features best predict the correctness of

model-generated reasoning, we train a neural network on both general features (CoT length, emotion volatility, and hedging rate) and with the 25 most consistently "harmful" non-lemmatized words (see Appendix 3, and combinations between the standard features and "harmful" words. We then test both its in-benchmark and cross-benchmark capability, by testing its performance on the benchmark it is trained on and the other benchmark.

For all four experiments, we employ a feed-forward neural network with two hidden layers (32 and 16 units, ReLU activations) and a sigmoid output. Models are trained using the Adam optimizer and binary cross-entropy loss. Given the strong class imbalance in the training data, particularly for HLE, whose accuracy for the minority class was only 8.9%, it is crucial to prevent the model from exploiting this imbalance with trivial predictions (e.g., always guessing "incorrect" for DeepSeek's answers on HLE would yield an accuracy of 91.4%). Consequently, we apply class weighting during neural network training, penalizing mistakes on the minority class more heavily, to encourage the model to learn non-trivial, meaningful patterns for both classes. All experiments are conducted on 30 seeds (0-29) to ensure robust estimation, with the results averaged across multiple runs. Training is performed on 80% of the original unbalanced data; we then balance the remaining 20% so that evaluation sets contain the same number of correct and incorrect examples, ensuring that performance consistently above 50% is indeed better than random guessing.

Figure 5 illustrates that adding additional features, such as CoT length or hedging rate, do not significantly enhance performance and, in some cases, even reduce cross-benchmark accuracy. This pattern is consistent across all random seeds and experimental setups. Notably, CoT length, which has been highlighted in previous studies (Ballon et al., 2025; Chen et al., 2024; Shojaee et al., 2025; Su et al., 2025; Wang et al., 2025), consistently provides the least predictive value and harms performance when included as an additional feature in all of our analyses. The ROC curves for each analysis can be found in Appendix C. These results suggest that this small set of words offers the most generalizable signal for predicting final response correctness, serving as lexical hints of uncertainty within CoT reasoning. The complete confusion matrices and Matthews Correlation Coefficient (MCC) for the best-performing neural network, which was trained using only these 25 words, are available in Appendix 5 and 6.

4 Discussion and Conclusion

Our results reveal a clear information hierarchy within CoT traces. Lexical uncertainty hints dominate: tokens such as *guess*, *stuck*, and *hard* reduce accuracy odds by up to 40 % relative to baseline (Fig.3). Intra-CoT sentiment volatility is weaker but complementary; a modest upward sentiment shift ($\Delta = 0.1$) coincides with peak accuracy on Omni-MATH, yet sentiment is uninformative on HLE (Fig.2). On the moderate difficulty Omni MATH benchmark we find a negative correlation between CoT length and answer accuracy, whereas no such link appears on the more demanding HLE set. This pattern matches earlier work reporting accuracy declines with longer CoTs on Omni MATH (Ballon et al., 2025) and the absence of this effect on highly challenging benchmarks (Shojaee et al., 2025; Opus & Lawsen, 2025).

This asymmetry can be exploited to obtain a practical heuristic: filter outputs whose CoT contains strong uncertainty words regardless of length or self-reported confidence. We found that even a small lexicon of "harmful" words drives performance on par with, or better than, more elaborate confidence-based schemes. Concretely, training a binary classifier on CoT length, emotion volatility, hedging rate plus the top 25 non-lemmatized "bad" words yields Matthews correlation coefficients (MCCs) of 0.229 when training and testing on HLE (and 0.239 cross-evaluated on Omni-Math), and 0.354 on Omni-Math (0.263 cross-evaluated on HLE). By contrast, simply thresholding self-reported confidence gives MCCs of only 0.085 (HLE) and 0.065 (Omni-Math). Strikingly, a lightweight rule, i.e. mark an answer wrong if any of the top 5 harmful words appears in its CoT, otherwise classify as correct, achieves MCC = 0.215 on HLE and 0.305 on Omni-Math. This suggests that a handful of carefully selected lexical cues can serve as a highly practical post-hoc filter. A further advantage of this technique is that it treats the CoT as plain text: no weight access, no second inference pass is necessary.

The contrast between DeepSeek-R1 and Claude 3.7 Sonnet illuminates the role of RLHF stylistics. Claude tends to generate longer, optimism-biased CoTs yet shares the same lexical uncertainty fingerprint as R1. This suggests that affective style might be an alignment artefact, whereas the full chain may not mirror

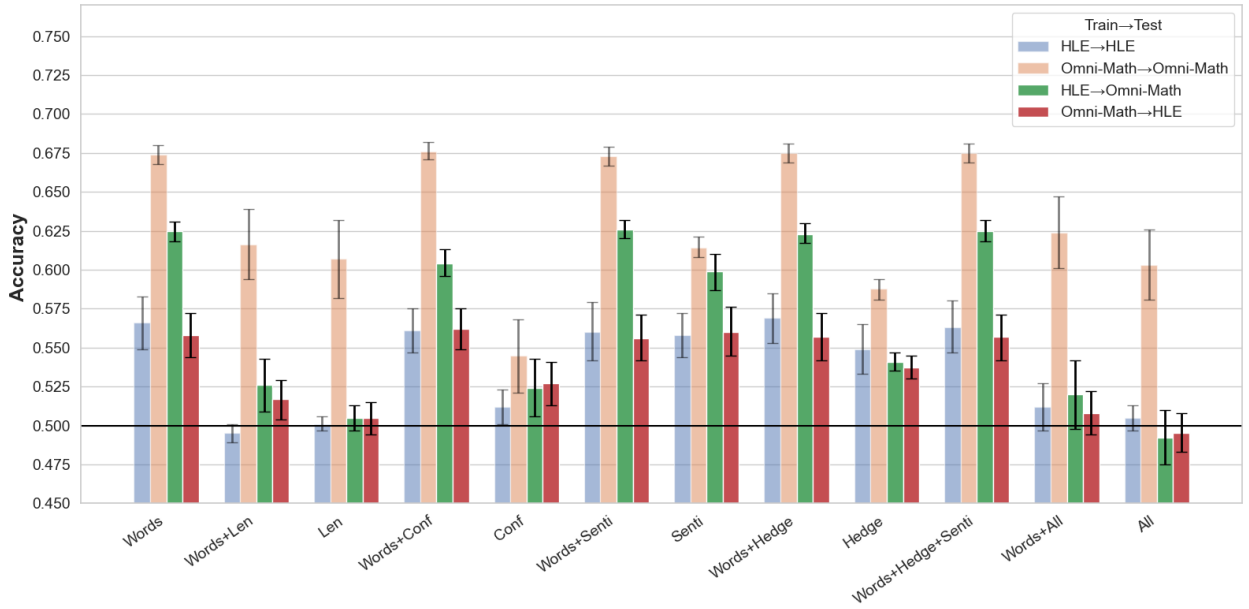


Figure 5: Each group of bars represents a different combination of input features: bag-of-words using 25 "harmful" words ("Words"), common surface features (CoT length, emotion, hedging), or their combinations. Neural nets were always trained on imbalanced data but evaluated on balanced test sets to ensure accuracy reflects true generalization rather than label bias. Results are averaged across 30 random seeds, with error bars showing variability. In-benchmark (blue, orange) performance is generally higher than out-of-benchmark (green, red). Adding additional emotion and hedging did not significantly improve cross-benchmark generalization, while adding length harmed both in-benchmark and cross-benchmark accuracy. This suggests that a small, targeted lexicon of harmful words serves as the most generalizable predictor of answer correctness across benchmarks.

the true multi-step search in latent space; the presence of hedging words does correlate with lower internal confidence (Chen et al., 2025b).

Limitations and future directions

In this paper, two English benchmarks were studied, so our findings might not hold up in languages with richer morphology and different hedging forms. Furthermore, our hedging list is fixed and English-focused, meaning we miss any new or non-English uncertainty phrases. The sentiment model we used (OpenAI o3-mini) brings its own calibration quirks. Using multiple raters or a human-scored baseline could smooth out existing noise. Additionally, our HLE grading depends on LLM–human agreement, which implies that the mislabels could hide subtle effects. Finally, correlation is not causation: those uncertainty markers could simply be more common around tougher, hidden problems in general.

Key avenues for future work include: (i) multilingual replication with adaptive hedging lexicons; (ii) adversarial prompting to force models to suppress hedging and test causal impact on accuracy; (iii) To understand RLHF's impact on uncertainty expression, future work could track changes in hedging frequency and harmful token usage between SFT and RLHF checkpoints. This includes testing whether RLHF amplifies or suppresses the lexical signals that predict incorrect reasoning.

Broader Impact Statement

This work shows that simple lexical cues in an LLM's Chain-of-Thought can help predict likely errors, offering a lightweight and model-agnostic way to improve calibration. Such signals could support safer deployment of reasoning models in high-stakes domains. Risks include adversarial evasion of these cues and false positives. Mitigation of these risks may involve combining our method with other calibration techniques and extending tests to multilingual and domain-diverse settings. Broader testing across languages and domains is needed to ensure robustness, prevent misuse, and identify effects that may differ outside the English reasoning benchmarks studied.

Data Availability

Data has been made available at: `https://figshare.com/s/7bb35af45e9fd7efe6cd`

Author Contributions

Acknowledgments

References

Anthropic. Claude 3.7 Sonnet, 2025a. URL `https://www.anthropic.com/news/claude-3-7-sonnet`. Large language model, version released 24 Feb 2025.

Anthropic. Claude's extended thinking, February 2025b. URL `https://www.anthropic.com/news/visible-extended-thinking`. Blog post.

David D Baek and Max Tegmark. Towards understanding distilled reasoning models: A representational approach. *arXiv preprint arXiv:2503.03730*, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Marthe Ballon, Andres Algaba, and Vincent Ginis. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*, 2025.

Sam Bowyer, Laurence Aitchison, and Desi R Ivanova. Position: Don't use the clt in llm evals with fewer than a few hundred datapoints. *arXiv preprint arXiv:2503.01747*, 2025.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025a.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? On the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025b.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.09714.

Cüneyt Demir. Hedging and academic writing: an analysis of lexical hedges. *Journal of language and linguistic studies*, 14(4):74–92, 2018.

Center for AI Safety. Humanity's last exam, March 2025. URL https://web.archive.org/web/20250302022138/https://agi.safe.ai/. Archived snapshot of the original page at https://agi.safe.ai/.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

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

Jason M Harley, Nigel Mantou Lou, Yang Liu, Maria Cutumisu, Lia M Daniels, Jacqueline P Leighton, and Lindsey Nadon. University students' negative emotions in a computer-based examination: the roles of trait test-emotion, prior test-taking methods and gender. *Assessment & Evaluation in Higher Education*, 46(6): 956–972, 2021.

Ken Hyland. Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, 18(3):349–382, 1998.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Qi Wang, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. *arXiv preprint arXiv:2505.22148*, 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *arXiv preprint arXiv:2504.09762*, 2025.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39-18, pp. 19260–19268, 2025.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

C. Opus and A. Lawsen. Comment on the illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.09250*, 2025.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Alexander Renkl. Learning from worked-out examples: A study on individual differences. *Cognitive science*, 21(1):1–29, 1997.

Paul A Schutz and Heather A Davis. Emotions and self-regulation during test taking. *Educational psychologist*, 35(4):243–256, 2000.

Paul A Schutz, Christine Distefano, Jeri Benson, and Heather A Davis. The emotional regulation during test-taking scale. *Anxiety, Stress & Coping*, 17(3):253–269, 2004.

Paul A Schutz, Jeri Benson, and Jessica T Decuir-Gunby. Approach/avoidance motives, test emotions, and emotional regulation related to testing. *Anxiety, Stress, & Coping*, 21(3):263–281, 2008.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL `https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf`.

Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *arXiv preprint arXiv:2505.13775*, 2025.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

Floriano Tori, Sara Tori, Imre Keseru, and Vincent Ginis. Performing sentiment analysis using natural language models for urban policymaking: An analysis of twitter data in brussels. *Data Science for Transportation*, 6(2):5, 2024.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. Thoughts are all over the place: On the underthinking of o1-like LLMs. *arXiv preprint arXiv:2501.18585*, 2025.

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

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Zixuan Wu, Francesca Lucchetti, Aleksander Boruch-Gruszecki, Jingmiao Zhao, Carolyn Jane Anderson, Joydeep Biswas, Federico Cassano, Molly Q Feldman, and Arjun Guha. Phd knowledge not required: A reasoning challenge for large language models. *arXiv preprint arXiv:2502.01584*, 2025.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. *arXiv preprint arXiv:2311.13240*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Accuracy Breakdown

Table 2: Accuracy (with count) for each model by HLE category, Omni-MATH tier, and Omni-MATH domain. For questions in Omni-MATH that span multiple domains, each domain instance is counted. Tiers are defined according to (Ballon et al., 2025). For both DeepSeek-R1 and Claude 3.7 Sonnet, the highest accuracy in each benchmark is shown in **bold**, and the second-highest is underlined. While both models excel in similar mathematical sub-disciplines on Omni-MATH, their areas of strength differ on HLE.

Dataset	Subset	Category/Tier/Domain	DeepSeek-R1 (n)	Claude 3.7 Sonnet (n)
HLE	Category	Engineering	**15%** (34)	6% (33)
		Math	9% (793)	**10%** (804)
		Biology/Medicine	7% (68)	6% (66)
		Computer Science/AI	6% (138)	6% (145)
		Humanities/Social Science	5% (95)	8% (90)
		Physics	5% (141)	6% (139)
		Other	3% (95)	2% (98)
		Chemistry	2% (63)	5% (63)
Omni-MATH	Tier	Tier 1	82% (1445)	80% (1445)
		Tier 2	70% (1304)	64% (1304)
		Tier 3	66% (649)	61% (649)
		Tier 4	67% (1030)	66% (1030)
Omni-MATH	Domain	Algebra	**78%** (2139)	**76%** (2139)
		Precalculus	76% (87)	75% (87)
		Other	75% (4)	75% (4)
		Number Theory	74% (916)	74% (916)
		Calculus	73% (128)	66% (128)
		Applied Mathematics	71% (805)	65% (805)
		Geometry	65% (1018)	60% (1018)
		Discrete Mathematics	58% (888)	55% (888)

B Booster Words

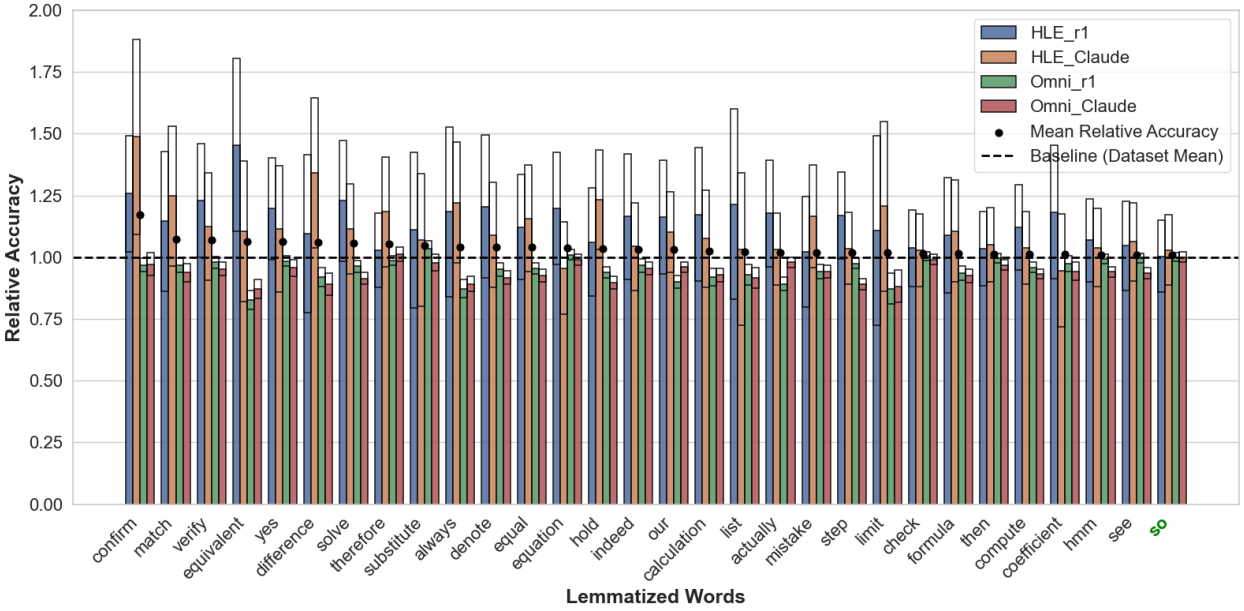
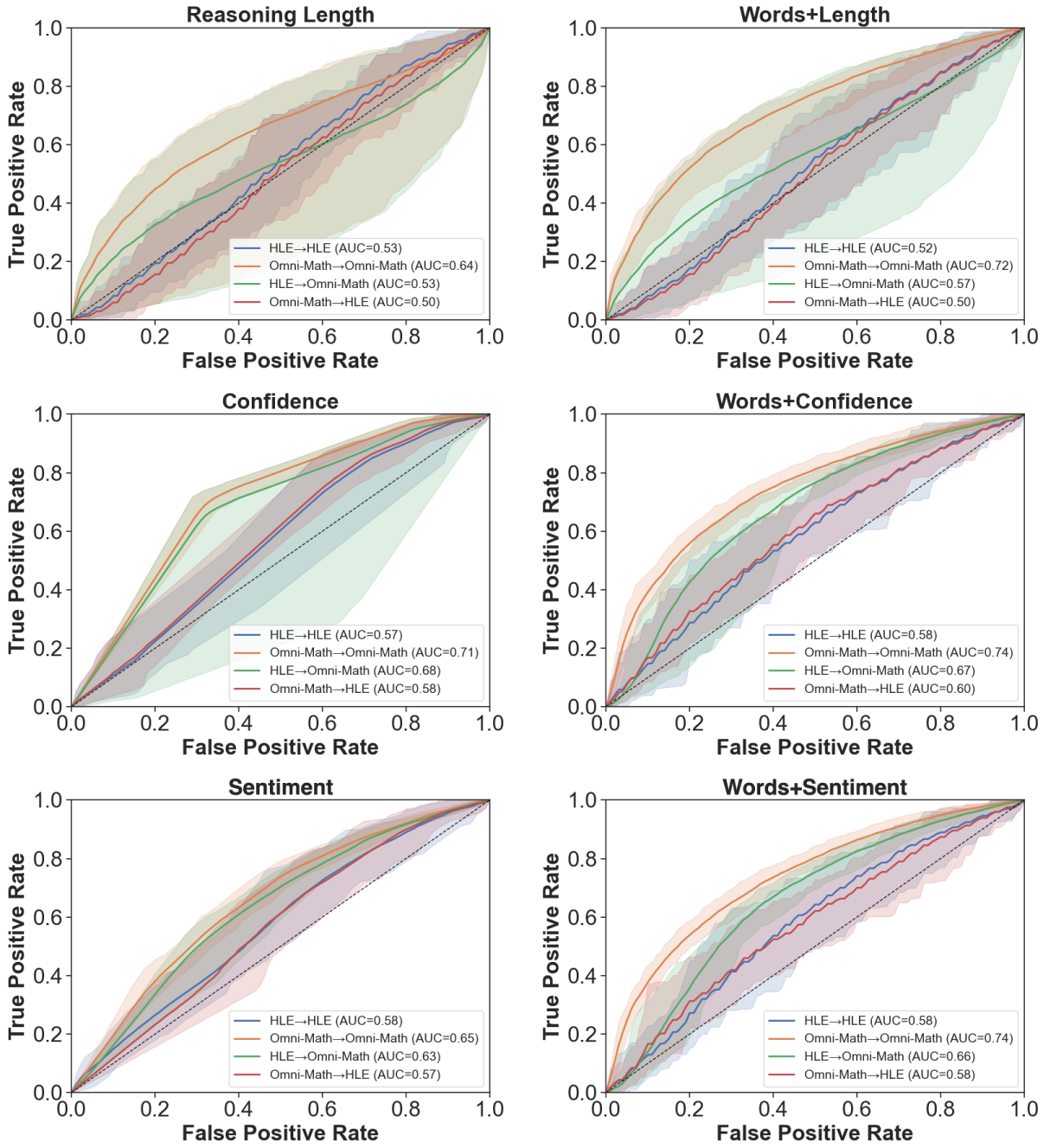


Figure 6: Relative accuracy of the 30 most beneficial lemmatized words across datasets and models. Words characteristic of mathematical reasoning (e.g., solve, therefore, equation) are most associated with higher accuracy. Only "so" (in green) appears universally positive, but this is likely due to chance.

C ROC curves


Reasoning Length
HLE→HLE (AUC=0.53)
Omni-Math→Omni-Math (AUC=0.64)
HLE→Omni-Math (AUC=0.53)
Omni-Math→HLE (AUC=0.50)
Words+Length
HLE→HLE (AUC=0.52)
Omni-Math→Omni-Math (AUC=0.72)
HLE→Omni-Math (AUC=0.57)
Omni-Math→HLE (AUC=0.50)
Confidence
HLE→HLE (AUC=0.57)
Omni-Math→Omni-Math (AUC=0.71)
HLE→Omni-Math (AUC=0.68)
Omni-Math→HLE (AUC=0.58)
Words+Confidence
HLE→HLE (AUC=0.58)
Omni-Math→Omni-Math (AUC=0.74)
HLE→Omni-Math (AUC=0.67)
Omni-Math→HLE (AUC=0.60)
Sentiment
HLE→HLE (AUC=0.58)
Omni-Math→Omni-Math (AUC=0.65)
HLE→Omni-Math (AUC=0.63)
Omni-Math→HLE (AUC=0.57)
Words+Sentiment
HLE→HLE (AUC=0.58)
Omni-Math→Omni-Math (AUC=0.74)
HLE→Omni-Math (AUC=0.66)
Omni-Math→HLE (AUC=0.58)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

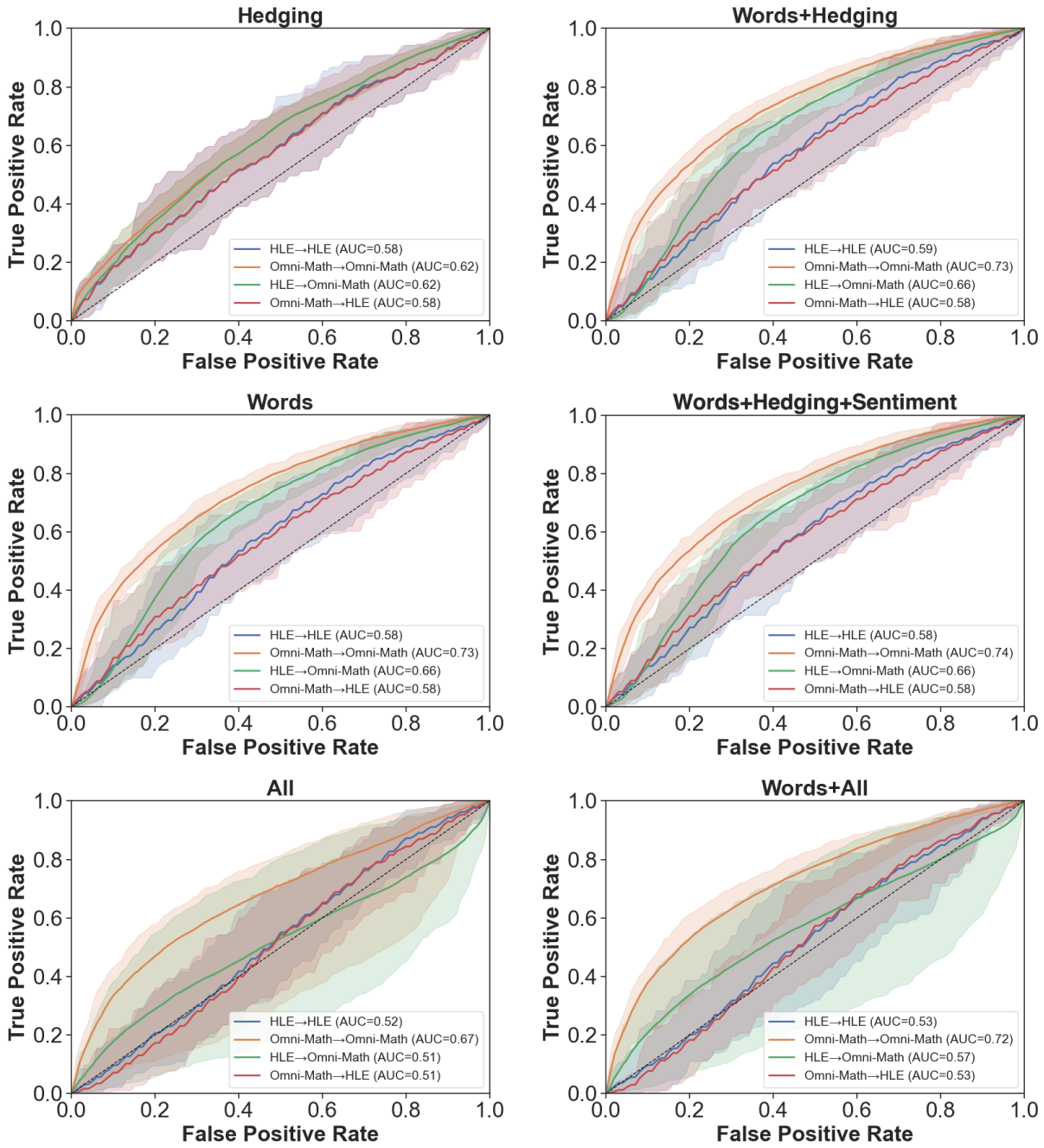


Figure 8: ROC curves across 12 experimental settings evaluating neural network classifiers for answer correctness. Each curve displays the mean true/false positive rates and 95% confidence interval (shaded region) across 30 random seeds. The models are trained and evaluated on balanced data for robust comparison. Area under the curve (AUC) values are reported in the legend. Models that use “harmful" words and abstain from using reasoning length perform the best and most consistently.

D Consistently Harmful Words Used in Training

Table 3: List of 25 non-lemmatized words most strongly and consistently associated with reduced accuracy ("harmful" words) across benchmarks.

Harmful Words for Accuracy				
complexity	guess	stuck	hard	involved
positions	involves	involve	information	comes
likely	probably	constraints	called	depend
missing	often	possibly	combination	based
four	require	structure	per	requires

E All Consistently Harmful Words

Table 4: List of all consistently "harmful" words and their 95% confidence interval for which the offset is given in (brackets). The 25 on average most "harmful" words were used as features for training the accuracy predicting neural network. As a consequence of rounding, some words have a relative accuracy equal to one.

Lemmatized Word	Mean Relative Accuracy	HLE-DeepSeek-R1	HLE-Claude	Omni-DeepSeek-R1	Omni-Claude
complexity	0.59 (0.09)	0.61 (0.29)	0.54 (0.24)	0.50 (0.07)	0.71 (0.08)
guess	0.67 (0.17)	0.69 (0.29)	0.76 (0.23)	0.42 (0.06)	0.80 (0.05)
stuck	0.70 (0.13)	0.71 (0.29)	0.81 (0.32)	0.51 (0.06)	0.78 (0.06)
hard	0.72 (0.12)	0.58 (0.25)	0.85 (0.24)	0.65 (0.06)	0.79 (0.04)
involve	0.74 (0.06)	0.77 (0.16)	0.65 (0.14)	0.76 (0.03)	0.77 (0.04)
information	0.75 (0.13)	0.57 (0.24)	0.88 (0.21)	0.75 (0.06)	0.82 (0.05)
come	0.76 (0.03)	0.76 (0.31)	0.77 (0.28)	0.80 (0.07)	0.72 (0.07)
likely	0.77 (0.11)	0.66 (0.18)	0.92 (0.19)	0.73 (0.04)	0.76 (0.04)
probably	0.77 (0.03)	0.81 (0.25)	0.76 (0.22)	0.78 (0.05)	0.75 (0.05)
position	0.79 (0.04)	0.74 (0.20)	0.82 (0.18)	0.80 (0.03)	0.79 (0.03)
constraint	0.79 (0.09)	0.68 (0.26)	0.82 (0.21)	0.76 (0.04)	0.89 (0.03)
call	0.79 (0.03)	0.75 (0.32)	0.80 (0.26)	0.80 (0.06)	0.83 (0.07)
miss	0.80 (0.10)	0.86 (0.27)	0.88 (0.20)	0.67 (0.05)	0.77 (0.04)
often	0.80 (0.03)	0.83 (0.26)	0.80 (0.22)	0.75 (0.06)	0.81 (0.08)
possibly	0.80 (0.05)	0.79 (0.29)	0.87 (0.24)	0.76 (0.06)	0.79 (0.05)
combination	0.80 (0.09)	0.78 (0.27)	0.70 (0.27)	0.90 (0.04)	0.84 (0.05)
base	0.81 (0.09)	0.76 (0.20)	0.93 (0.16)	0.72 (0.04)	0.83 (0.03)
structure	0.81 (0.07)	0.78 (0.20)	0.92 (0.20)	0.78 (0.04)	0.77 (0.04)
per	0.81 (0.07)	0.90 (0.24)	0.73 (0.27)	0.81 (0.05)	0.81 (0.07)
depend	0.82 (0.03)	0.84 (0.14)	0.86 (0.14)	0.79 (0.03)	0.79 (0.03)
complex	0.82 (0.09)	0.76 (0.22)	0.94 (0.18)	0.74 (0.04)	0.84 (0.03)
etc	0.82 (0.07)	0.88 (0.21)	0.72 (0.23)	0.86 (0.03)	0.83 (0.04)
very	0.82 (0.07)	0.76 (0.23)	0.90 (0.17)	0.78 (0.04)	0.86 (0.03)
beyond	0.82 (0.04)	0.88 (0.33)	0.79 (0.32)	0.81 (0.05)	0.81 (0.06)
help	0.82 (0.05)	0.87 (0.26)	0.75 (0.27)	0.87 (0.03)	0.82 (0.05)
direction	0.83 (0.05)	0.87 (0.31)	0.87 (0.25)	0.77 (0.05)	0.79 (0.04)
possibility	0.83 (0.10)	0.70 (0.29)	0.81 (0.24)	0.91 (0.06)	0.89 (0.05)
else	0.83 (0.05)	0.88 (0.27)	0.79 (0.26)	0.85 (0.04)	0.80 (0.05)
actual	0.83 (0.10)	0.94 (0.25)	0.89 (0.24)	0.80 (0.04)	0.71 (0.05)
mention	0.83 (0.12)	0.89 (0.32)	0.97 (0.24)	0.72 (0.07)	0.75 (0.05)
group	0.84 (0.12)	0.66 (0.26)	0.94 (0.30)	0.88 (0.06)	0.86 (0.06)
without	0.84 (0.06)	0.85 (0.18)	0.92 (0.17)	0.78 (0.03)	0.80 (0.03)
require	0.84 (0.06)	0.81 (0.12)	0.92 (0.16)	0.79 (0.02)	0.83 (0.03)
far	0.84 (0.06)	0.80 (0.31)	0.90 (0.18)	0.77 (0.05)	0.88 (0.03)
additional	0.84 (0.06)	0.88 (0.32)	0.90 (0.23)	0.75 (0.06)	0.83 (0.04)
space	0.84 (0.10)	0.98 (0.29)	0.86 (0.23)	0.78 (0.07)	0.75 (0.07)

Table continues on the next two pages

Lemmatized Word	Mean Relative Accuracy	HLE-DeepSeek-R1	HLE-Claude	Omni-DeepSeek-R1	Omni-Claude
system	0.84 (0.02)	0.85 (0.27)	0.81 (0.23)	0.85 (0.04)	0.86 (0.04)
too	0.84 (0.08)	0.75 (0.18)	0.95 (0.19)	0.81 (0.03)	0.85 (0.03)
certain	0.85 (0.07)	0.91 (0.19)	0.90 (0.19)	0.77 (0.03)	0.81 (0.04)
lead	0.85 (0.06)	0.78 (0.13)	0.91 (0.15)	0.86 (0.02)	0.85 (0.03)
part	0.86 (0.04)	0.82 (0.15)	0.89 (0.16)	0.88 (0.03)	0.82 (0.03)
unless	0.86 (0.02)	0.87 (0.21)	0.83 (0.28)	0.85 (0.03)	0.87 (0.05)
specific	0.86 (0.03)	0.88 (0.19)	0.88 (0.15)	0.81 (0.03)	0.86 (0.03)
proceed	0.86 (0.12)	1.00 (0.26)	0.88 (0.32)	0.70 (0.05)	0.86 (0.06)
change	0.86 (0.08)	0.88 (0.22)	0.96 (0.18)	0.78 (0.04)	0.82 (0.03)
common	0.86 (0.09)	0.94 (0.26)	0.74 (0.21)	0.86 (0.04)	0.91 (0.04)
pattern	0.86 (0.06)	0.78 (0.30)	0.92 (0.22)	0.85 (0.03)	0.89 (0.03)
high	0.86 (0.09)	0.94 (0.28)	0.94 (0.28)	0.80 (0.05)	0.77 (0.07)
suggest	0.86 (0.09)	0.81 (0.20)	1.00 (0.18)	0.79 (0.03)	0.86 (0.04)
angle	0.86 (0.05)	0.90 (0.25)	0.90 (0.19)	0.81 (0.04)	0.84 (0.03)
confuse	0.86 (0.10)	0.92 (0.25)	0.98 (0.25)	0.77 (0.05)	0.79 (0.05)
assumption	0.87 (0.07)	0.92 (0.26)	0.90 (0.27)	0.77 (0.04)	0.90 (0.05)
analysis	0.87 (0.06)	0.94 (0.33)	0.87 (0.21)	0.79 (0.05)	0.88 (0.04)
complicated	0.87 (0.03)	0.92 (0.19)	0.84 (0.22)	0.87 (0.03)	0.86 (0.03)
similar	0.87 (0.08)	0.97 (0.20)	0.91 (0.21)	0.79 (0.03)	0.82 (0.04)
multiple	0.88 (0.04)	0.91 (0.23)	0.83 (0.20)	0.86 (0.03)	0.90 (0.03)
much	0.88 (0.07)	0.90 (0.26)	0.96 (0.20)	0.82 (0.04)	0.83 (0.03)
specifically	0.88 (0.04)	0.93 (0.28)	0.89 (0.17)	0.85 (0.05)	0.85 (0.03)
relate	0.88 (0.07)	0.92 (0.17)	0.96 (0.17)	0.84 (0.03)	0.80 (0.03)
next	0.88 (0.05)	0.91 (0.23)	0.81 (0.22)	0.93 (0.03)	0.87 (0.03)
count	0.88 (0.04)	0.89 (0.28)	0.94 (0.26)	0.87 (0.04)	0.84 (0.04)
new	0.89 (0.07)	0.90 (0.30)	0.98 (0.23)	0.83 (0.04)	0.83 (0.04)
relation	0.89 (0.03)	0.85 (0.31)	0.93 (0.25)	0.87 (0.05)	0.89 (0.04)
assume	0.89 (0.07)	0.95 (0.15)	0.93 (0.14)	0.80 (0.03)	0.87 (0.02)
simple	0.89 (0.07)	0.97 (0.25)	0.92 (0.14)	0.81 (0.04)	0.84 (0.02)
third	0.89 (0.07)	0.80 (0.25)	0.97 (0.25)	0.90 (0.03)	0.88 (0.03)
look	0.89 (0.07)	0.83 (0.16)	0.98 (0.12)	0.84 (0.03)	0.90 (0.02)
thing	0.89 (0.06)	0.96 (0.26)	0.92 (0.22)	0.82 (0.04)	0.86 (0.04)
within	0.90 (0.08)	0.95 (0.30)	0.97 (0.24)	0.82 (0.04)	0.83 (0.05)
correspond	0.90 (0.07)	0.94 (0.17)	0.97 (0.15)	0.82 (0.03)	0.85 (0.03)
lower	0.90 (0.10)	0.99 (0.26)	0.98 (0.29)	0.82 (0.04)	0.80 (0.06)
valid	0.90 (0.07)	0.80 (0.27)	0.97 (0.23)	0.90 (0.03)	0.91 (0.03)

Table continues on the next page

Lemmatized Word	Mean Relative Accuracy	HLE-DeepSeek-R1	HLE-Claude	Omni-DeepSeek-R1	Omni-Claude
single	0.90 (0.10)	0.99 (0.23)	0.98 (0.21)	0.81 (0.04)	0.81 (0.04)
pair	0.90 (0.03)	0.88 (0.23)	0.95 (0.21)	0.89 (0.03)	0.88 (0.03)
form	0.90 (0.05)	0.85 (0.20)	0.97 (0.17)	0.88 (0.03)	0.92 (0.03)
something	0.90 (0.05)	0.95 (0.19)	0.94 (0.17)	0.88 (0.03)	0.85 (0.03)
itself	0.91 (0.05)	0.94 (0.25)	0.95 (0.22)	0.86 (0.04)	0.86 (0.04)
those	0.91 (0.06)	0.97 (0.20)	0.91 (0.20)	0.91 (0.03)	0.84 (0.03)
distinct	0.91 (0.08)	0.99 (0.35)	0.96 (0.27)	0.86 (0.04)	0.83 (0.04)
general	0.91 (0.03)	0.94 (0.24)	0.93 (0.18)	0.89 (0.04)	0.89 (0.03)
don	0.91 (0.03)	0.95 (0.20)	0.92 (0.15)	0.89 (0.03)	0.89 (0.02)
integer	0.91 (0.05)	0.86 (0.21)	0.89 (0.19)	0.93 (0.02)	0.97 (0.02)
why	0.91 (0.09)	0.99 (0.36)	1.00 (0.24)	0.86 (0.05)	0.81 (0.04)
even	0.91 (0.03)	0.95 (0.18)	0.90 (0.16)	0.91 (0.02)	0.89 (0.03)
could	0.91 (0.05)	0.93 (0.17)	0.98 (0.16)	0.87 (0.03)	0.88 (0.03)
their	0.91 (0.06)	0.99 (0.18)	0.89 (0.18)	0.91 (0.03)	0.86 (0.03)
still	0.92 (0.06)	0.91 (0.18)	0.99 (0.15)	0.86 (0.03)	0.90 (0.02)
add	0.92 (0.04)	0.91 (0.19)	0.96 (0.20)	0.95 (0.02)	0.86 (0.03)
linear	0.92 (0.05)	0.99 (0.28)	0.89 (0.27)	0.90 (0.04)	0.90 (0.05)
factor	0.92 (0.08)	0.95 (0.19)	0.80 (0.18)	0.98 (0.02)	0.96 (0.03)
consider	0.92 (0.05)	0.97 (0.12)	0.96 (0.11)	0.88 (0.02)	0.88 (0.02)
figure	0.93 (0.07)	0.94 (0.18)	0.99 (0.25)	0.94 (0.03)	0.83 (0.04)
through	0.93 (0.06)	0.97 (0.23)	0.99 (0.17)	0.86 (0.03)	0.88 (0.03)
such	0.93 (0.02)	0.90 (0.16)	0.95 (0.16)	0.92 (0.02)	0.93 (0.02)
question	0.93 (0.07)	0.98 (0.17)	0.99 (0.16)	0.90 (0.03)	0.85 (0.03)
however	0.93 (0.07)	0.99 (0.15)	0.99 (0.15)	0.88 (0.02)	0.87 (0.03)
start	0.93 (0.05)	0.95 (0.14)	0.99 (0.13)	0.93 (0.02)	0.86 (0.02)
like	0.93 (0.03)	0.97 (0.16)	0.95 (0.15)	0.91 (0.02)	0.90 (0.02)
work	0.93 (0.03)	0.90 (0.18)	0.97 (0.15)	0.92 (0.02)	0.94 (0.02)
positive	0.94 (0.03)	0.93 (0.25)	0.89 (0.22)	0.96 (0.03)	0.96 (0.03)
express	0.94 (0.03)	0.96 (0.20)	0.97 (0.19)	0.90 (0.03)	0.92 (0.03)
example	0.94 (0.05)	0.96 (0.15)	1.00 (0.20)	0.92 (0.02)	0.88 (0.03)
they	0.94 (0.04)	0.99 (0.12)	0.97 (0.12)	0.92 (0.02)	0.90 (0.02)
my	0.95 (0.07)	0.99 (0.23)	1.00 (0.14)	0.85 (0.04)	0.94 (0.02)
up	0.95 (0.04)	0.98 (0.17)	0.97 (0.16)	0.94 (0.02)	0.89 (0.03)
take	0.95 (0.05)	0.99 (0.16)	1.00 (0.14)	0.93 (0.02)	0.88 (0.02)
down	0.95 (0.07)	0.98 (0.23)	1.00 (0.19)	0.97 (0.03)	0.85 (0.03)
need	0.96 (0.02)	0.97 (0.12)	0.98 (0.11)	0.94 (0.02)	0.94 (0.02)
solution	0.96 (0.04)	1.00 (0.30)	0.99 (0.23)	0.92 (0.03)	0.94 (0.03)
on	0.96 (0.03)	0.99 (0.15)	0.98 (0.14)	0.93 (0.02)	0.94 (0.02)
an	0.96 (0.04)	0.99 (0.16)	0.99 (0.14)	0.92 (0.02)	0.95 (0.02)
be	0.97 (0.02)	0.98 (0.07)	0.99 (0.06)	0.95 (0.01)	0.94 (0.01)
not	0.97 (0.02)	0.98 (0.15)	1.00 (0.14)	0.94 (0.02)	0.96 (0.02)

Table 5: Confusion matrices and Matthews Correlation Coefficients (MCC) for our best-performing neural network, trained using only the 25 most consistently harmful words as input features. Evaluations were conducted on a balanced test set, with results shown for a single random seed (seed = 2).

Train→Test	**Actual**	**Pred. 0**	**Pred. 1**	**MCC**
HLE→HLE	0	40	30	0.229
	1	24	46	
HLE→Omni-Math	0	241	276	0.239
	1	123	394	
Omni-Math→Omni-Math	0	355	162	0.354
	1	172	345	
Omni-Math→HLE	0	57	13	0.263
	1	40	30	

Table 6: Confusion matrices and Matthews Correlation Coefficients (MCC) for two heuristic methods evaluated on the HLE and Omni-Math datasets using a balanced test set. The 5 Harmful Words heuristic predicts failure if any of the five most consistently harmful words appear in the reasoning text. The Confidence heuristic uses self-reported confidence to probabilistically predict correctness by flipping coin weighted by its reported confidence level, to assign a grade of either 0 or 1. Our results illustrate that the word-based heuristic consistently outperforms the probabilistic confidence-based method.

Heuristic	**Dataset**	**Actual**	**Pred. 0**	**Pred. 1**	**MCC**
5 Most Harmful Words	HLE	0	46	24	0.215
		1	31	39	
	Omni-Math	0	317	200	0.305
		1	160	357	
Confidence	HLE	0	1	69	0.085
		1	0	70	
	Omni-Math	0	35	482	0.065
		1	20	497	