Effort and Confidence: Surface Markers in Reasoning Traces Predict Correctness on Insight Puzzles

Richard A. J. Mann
City St George's,
University of London
Richard.Mann.2@citystgeorges.ac.uk

Pranava Madhyastha
City St George's, University of London
The Alan Turing Institute
Pranava.Madhyastha@citystgeorges.ac.uk

Abstract

Multimodal large language models with *reasoning* abilities produce extended thinking traces before committing to an answer. Whether these natural language self-descriptions of the models' problem-solving behaviours faithfully reflect their internal uncertainty, and whether that uncertainty affects the committed output, are open questions with significant implications for how we evaluate and trust model reasoning processes. We present a focused study that investigates this relationship using insight puzzles, which require a non-linear reinterpretation of semantic, phonetic or visual clues and thereby produce a discrete resolution event with a characteristic surface signature. Our study reveals three findings. First, the correctness signal in *reasoning* traces across four insight domains is largely recoverable from surface effort markers, where a simple six-feature keyword-based classifier matches or outperforms the richer semantic models we tested, predicting correctness at average precision 0.85 against a 0.62 base rate and rising to 0.89 once an additional trace-derived confidence feature is added. Second, this signal is only partly shared with mathematical reasoning. Length and duration effort markers predict correctness on math traces too, but the hedging and resolution language that distinguishes insight traces is largely absent from math, where models express difficulty through trace length rather than uncertainty and confidence markers. Third, the model's self-reported confidence is systematically miscalibrated relative to the signal present in its own trace, with an external classifier reading the same text roughly five times better calibrated.

1 Introduction

The ability to reason[1] before generating the final answer is now a standard capability of most frontier multimodal language models. These models produce extended articulation of thinking traces, sometimes spanning thousands of tokens, before committing to an answer, and there is growing evidence that the quality of this reasoning process is consequential for downstream performance (Marjanović et al., 2025). However, the relationship between what a model reasons and what it ultimately outputs is still poorly understood. Whether a model's trace faithfully reflects its uncertainty, and whether that uncertainty survives the compression into a committed answer, are questions that current evaluation practices do not directly address.

We adopt insight puzzles across four domains as our testbed (Figure 1 shows one example of each). Insight puzzles (see Metcalfe and Wiebe, 1987, for a comprehensive discussion) are problems whose surface presentation under-determines the solution method. Solving requires a non-obvious reinterpretation of semantic, phonetic, or visual signals rather than execution of a standard procedure. Previous work labels these using terms such as lateral thinking (Lee et al., 2025). As math benchmarks saturate (Balunović et al., 2025), insight puzzles offer a complementary testbed where the trace signatures of problem-solving differ structurally from methodical mathematical execution.

We present a focused study on seven frontier VLMs spanning four insight puzzle domains and two math-based benchmarks. Our analysis reveals three findings that together characterise the trace-output relationship in insight reasoning. First, the correctness signal is largely recoverable at the surface. A simple keyword-based linear model matches or outperforms more sophisticated alternatives with richer semantic representations. Second, this signal is only partly shared with mathematical reasoning. Length and duration effort markers still predict correctness on math traces, but the hedging and resolution language that carries much of the insight signal is largely absent from math. This aligns with Zhao et al. (2025), who find that math aha-language is largely decorative, a conclusion

[1] explicit chain-of-thought rendered as *thinking traces*

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 696–717
October 17-21, 2026 ©2026 Association for Computational Linguistics

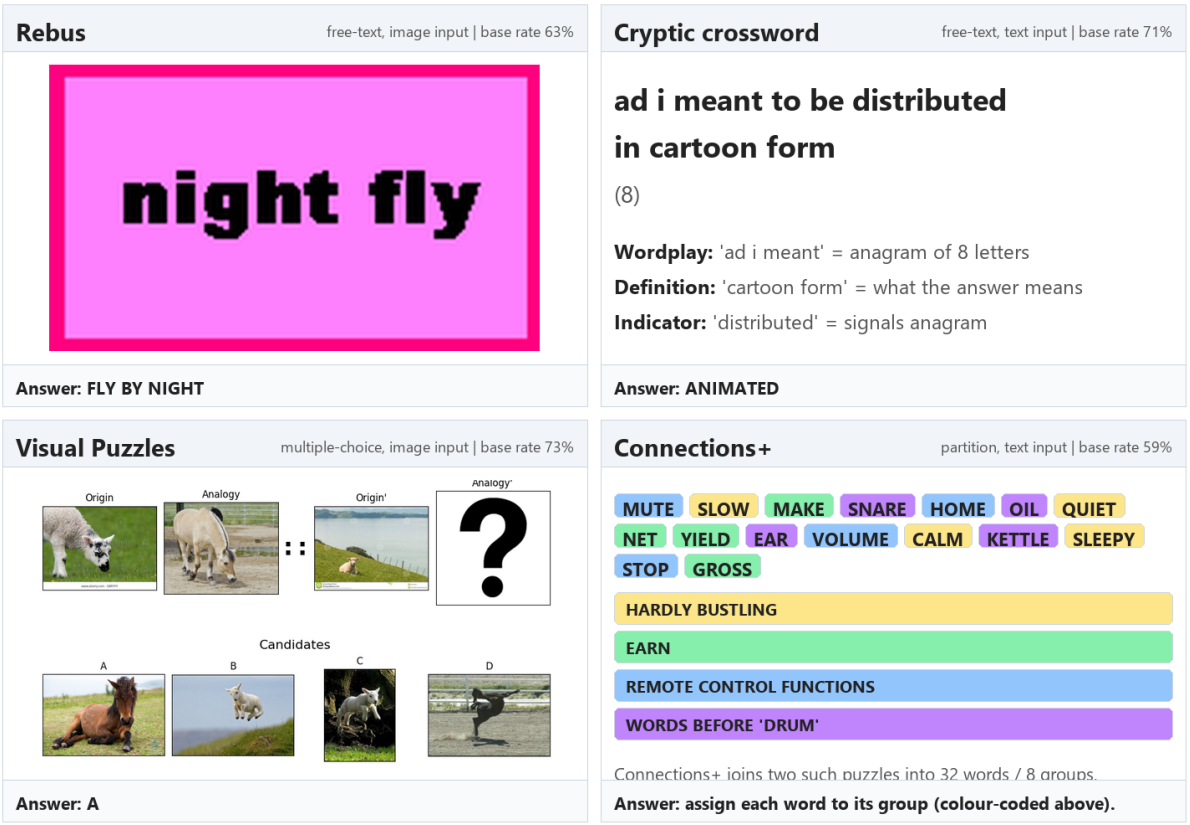


Figure 1: Here we show one puzzle from each of the four insight domains, with the seven-model cohort base rate in each panel header. All four need a non-obvious reinterpretation of the clue rather than a standard procedure, but they differ in what the model is shown and in what counts as an answer. **Rebus** (free-text, image input): "FLY BY NIGHT" from "night fly". **Cryptic crossword** (free-text, text input): "ANIMATED" is the anagram of "ad i meant" (definition "cartoon form"). **Visual Puzzles** (multiple-choice, image input): visual analogy with correct option A. **Connections+** (partition, text input): 16 words from one NYT Connections puzzle, pills colour-coded by their 4-group partition. Connections+ joins two such puzzles into 32 words and 8 groups.

we reach by a different, feature-level route. Third, the model's self-reported confidence in its committed output is systematically less calibrated than the information present in its own *reasoning trace*. An external classifier reading the trace is roughly five times better calibrated than the model's self-reported confidence score, indicating that models are predominantly overconfident.

2 Related work

The closest prior work is Pawitan and Holmes (2025), who assess confidence in LLM chain-of-thought reasoning via persistence under prompting and self-reported confidence on text-only causal-judgement and probability puzzles. We extend this framing to VLMs and to insight puzzles. DeepSeek-AI (2025) coined the modern usage of aha moments as emergent training-time phenomena, and Meng et al. (2025) showed visual aha exists qualitatively in VLMs on math diagrams. The field has since treated aha as unreliable signal to be controlled (Hu et al., 2025; Ha et al., 2025), with Zhao et al. (2025) finding that aha steps in math traces are often decorative under counterfactual intervention. Our claim takes the opposite position in a different domain. Insight puzzles depend on a non-obvious reinterpretation and the aha moment is a critical part of the solution, not cosmetic. On the relationship between effort and correctness, Marjanović et al. (2025) document rumination in math reasoning traces and find effort and correctness are related, which motivates our feature design. We show the same relationship holds across insight domains and transfers cross-domain with six surface features. Finally, Kadavath et al. (2022) claim that models mostly know what they know. Our self-report miscalibration finding refines this, as introspection in the trace does not always survive compression into the committed output. Most directly related, Pandey

et al. (2026) use a hedge-to-verification ratio in reasoning traces as a single-pass uncertainty signal on general and mathematical benchmarks. We extend this to insight puzzles (including visual ones), add a local aha-confidence feature that lifts prediction beyond hedge and effort markers, demonstrate cross-domain transfer, and surface a self-report calibration gap. Vanhoyweghen et al. (2025) likewise find lexical and length cues predictive of accuracy in reasoning chains. We add cross-domain and cross-model transfer, a discrete-resolution mechanism, and the insight-versus-math boundary. On effort and difficulty, Su et al. (2025) show that models miscalibrate reasoning length to problem difficulty. We build on this to argue that predicting difficulty from the trace is itself useful, since difficulty is unobservable at solve time, and that our classifier is therefore not simply a difficulty proxy because it significantly outperforms a peer-difficulty oracle. Hedging is a long-studied epistemic device in human language (Lakoff, 1973). We show these markers recur in machine reasoning traces and track the model's genuine token-level uncertainty (Appendix G), rather than a stylistic artefact.

3 Methodology

Our focus is to evaluate if the tendency of the model to generate correct outputs can be determined from the surface of a reasoning trace, without access to the final answer generated by the model. We use the term 'surface' to mean what is readable from trace generation without interpreting reasoning, such as its length, its timing and its lexical markers. Two of the six markers are not textual, since elapsed time and the provider-reported token count are process measures, and partly reflect provider behaviour as well as model behaviour.

Features. Motivated by the finding in Marjanović et al. (2025) that effort and correctness of thinking traces are related, we derive six features that capture how hard the model worked on a problem. Three measure raw effort: the API-reported reasoning-token count, elapsed wall-clock time, and trace character count. Another three capture hedge language distribution within the trace, i.e., epistemic markers such as *maybe*, *I think*, *not sure* (Farkas et al., 2010), via their per-word rate, their ratio to total tokens, and the positional variance of their occurrence. Together these form an effort signal, with an assumption that harder problems elicit more tokens, more time, and more hedging, and harder problems are more likely to be wrong. These six are a strong subset of a more comprehensive 21-feature length-and-rate baseline (Table 1). The remaining features either duplicate this signal or do not transfer consistently across domains. No single feature carries the classifier. In a leave-one-out ablation, removing any one of the six changes mean average precision (AP) by at most 0.03, and no feature alone matches the six together (Appendix H). The set is robust because different features carry the signal in different domains and models, not because any one is decisive.

Marker selection. The hedge and commitment lists are trace-derived. They were drawn up from a review of a 20-puzzle Rebus sample (134 traces) and confirmed on review as logical epistemic markers. Three hedge and eight commitment patterns were added later from an n-gram pass over the 100-puzzle Rebus set. Features derived from data should be tested on data that played no part in deriving them, and the collection order gives us that. Both lists were frozen before the Cryptic, VisualPuzzles and Connections+ corpora were collected, so 300 of the 400 puzzles took no part in their construction and the leave-one-domain-out results on those three domains are unaffected by it. Within Rebus, where the 20-puzzle sample does fall inside the evaluation set, dropping those puzzles from training and evaluation does not weaken the classifier. Lift over base rate is $+18.0$pp (percentage points) on the remaining 80, against $+16.9$pp on all 100 (further details are provided in Appendix I).

A seventh feature targets a qualitatively different signal in the form of the model's trace commitment language when it considers the answer it ultimately chooses. Insight puzzles, in most cases, characteristically produce a discrete resolution event, referred to as an *aha moment* (Metcalfe and Wiebe, 1987). Given the way the LLMs are post-trained, we expect this to leave a detectable surface signature. We operationalise this with an LLM judge (Gemini 3 Flash) that reads the thinking trace alongside the set of candidate answers and returns a per-candidate confidence score on a 0–100 scale anchored to commitment-language cues (*that's it!* high, *I give up* low; validated against a human judge and two LLM judges, Appendix D; Connections+ in Appendix B). Unlike the six effort features, this feature needs the model's answer, though never the ground truth, so that the judge can score the

Family / Feature	Description
Length	
thinking_char_count	Trace characters
thinking_word_count	Trace words
elapsed	Thinking time (s)
tokens_thinking_proxy	Unified token count
Trajectory	
n_candidate_mentions	Distinct candidates
candidate_churn	Candidate switches
Per-word rates	
hedge_rate	Hedges / word
confidence_rate	Commitment / word
explicit_rejection_rate	Rejections / word
candidate_switch_rate	Switches / word
revision_rate	Revisions / word
self_correction_rate	Self-corr. / word
question_mark_rate	"?" / word
exclamation_mark_rate	"!" / word
Normalised text	
hedge_ratio	Hedge / total
hedge_shift	1st/2nd-half hedge
hedge_density_last_third	Final-third hedge
hedge_position_variance	Hedge-position var.
bigram_repetition_rate	Repeated bigrams
trigram_repetition_rate	Repeated trigrams
unique_trigram_ratio	Unique trigrams

Table 1: The 21-feature LEN+RATES baseline. The 6-feature deployed classifier uses the subset in bold: tokens_thinking_proxy, elapsed, thinking_char_count, hedge_rate, hedge_ratio, hedge_position_variance.

candidate the model committed to.

Alternative representations. We also evaluated richer semantic alternatives: frozen sentence embeddings (MiniLM-L6 (Wang et al., 2020), bge-m3 (Chen et al., 2024)) and a fine-tuned ModernBERT classifier (Warner et al., 2024).

Classifier. We combine the effort and confidence features in a two-stage stacked logistic regression. The first stage fits a logistic regression on the six z-scored effort features (with an L2 regularisation penalty). The second stage fits a logistic regression on that predicted probability together with the normalised confidence score. We chose logistic regression deliberately because its linearity in the log-odds keeps each feature's contribution interpretable and makes its behaviour on an unseen domain or model easy to inspect.

Evaluation protocol. We evaluate under three nested generalisation conditions. To test the performance across domains, we use leave-one-domain-out (LODO) strategy, where we train on three insight domains and test on the fourth. Within-domain performance uses leave-one-puzzle-out cross-validation (LOPO, 10-fold), which measures how well the classifier transfers across puzzles of the same type. Cross-model performance uses leave-one-model-out (LOMO), which tests whether the signal generalises across model families. We report average precision (AP) as the primary metric, with 95% confidence intervals from a puzzle-clustered bootstrap ($N = 500$).

4 Experimental setup

Models. We evaluate seven frontier reasoning VLMs: Gemini 2.5 Pro, Gemini 3 Pro, Gemini 3 Flash[2], GPT-5, Claude Opus 4.6, Claude Sonnet 4.6, and Qwen3-VL-235B Thinking (more details in Appendix B on provider-specific reporting).

Insight corpus. Three of the four insight domains are established public benchmarks. The fourth, Connections+, is constructed for this work from public NYT Connections puzzles. The primary corpus comprises 400 puzzles distributed across four domains: Rebus (Das et al., 2025) (image-based, free-text answer), Cryptic crosswords sampled from Cryptonite (Efrat et al., 2021) (text-based, free-text answer), Visual Puzzles (Song et al., 2025) (image-based, multiple-choice), and Connections+ (text-based, partition task; constructed for this work by pairing two NYT Connections daily puzzles (Loredo Lopez et al., 2025) into a 32-word grid; see Appendix A). Each puzzle is attempted by all seven models (six on Connections+, where GPT-5 repeatedly exhausted its token budget before emitting an answer) on two passes: an un-augmented pass with a standard answer prompt, and an augmented pass that additionally requests an explicit 0–100 self-confidence score (Per-domain counts and data sources are in Appendix A; per-task statistics in Appendix K; one example per domain in Figure 1). Unlike our math corpus, these insight benchmarks are not post-cutoff, but their low base rates argue against contamination driving the result (see Limitations).

Math corpus. To test if the insight-trace features transfer to a structurally different reasoning domain, we use two math datasets: (i) a **post-cutoff** set of 50 problems in total drawn from HMMT (November 2025 and February 2026) and AIME 2026 (Balunović et al., 2025), all released after the confirmed training cutoff of every model in the cohort; and (ii) a **pre-cutoff contamination reference** of 100 problems sampled from MATH-500 (Lightman et al., 2024; Hendrycks et al., 2021) for the saturation comparison reported in Appendix F.

[2]We use Gemini in more than one role, and set out the basis for this in the Limitations.

Held-out domain	Model acc.	Effort AP	+conf AP
Rebus	0.60	0.82	0.85
Cryptic	0.65	0.92	0.96
VP	0.67	0.83	0.87
Connections+	0.54	0.84	0.88
Mean	**0.62**	**0.85**	**0.89**

Table 2: LODO AP for per-trace correctness across four insight domains (train on three, test on held-out). **Model acc.**: 7-model cohort mean on the domain, and the base rate the AP columns should be read against. **Effort AP**: 6-feature classifier. **+conf AP**: effort + LLM-judge confidence, stacked LR. Full LOMO breakdown and 95% CIs in Appendix C.

Held-out model	Base	Effort AP	+conf AP
Gemini 3 Pro	0.89	0.940	0.943
Gemini 3 Flash	0.82	0.911	0.925
GPT-5	0.71	0.890	0.915
Claude Opus 4.6	0.76	0.815	0.846
Claude Sonnet 4.6	0.67	0.770	0.813
Qwen3-VL-235B	0.30	0.488	0.573
Gemini 2.5 Pro	0.59	0.739	0.859
Mean	**0.68**	**0.793**	**0.839**

Table 3: Cross-model (LOMO) transfer: train on six models, test on the held-out seventh (un-augmented pass). **Base**: the held-out model's accuracy, which is the base rate its AP should be read against. **Effort AP**: 6-feature classifier. **+conf AP**: effort + LLM-judge confidence. Per-fold lift and 95% CIs in Appendix C.

The post-cutoff constraint is important, as math benchmark saturation is partly a contamination artefact, and we want a setting that challenges the models and does not contain memorisation traces.

5 Results

5.1 The effort signal predicts correctness

Here we investigate whether a thinking trace, read without knowledge of the answer, carries enough signal to predict whether the model got the puzzle right. Across four insight domains, the answer is yes. The 6-feature effort classifier achieves within-domain (LOPO) AP of 0.81–0.91, predicting per-trace correctness from token count, elapsed time, character count, and three hedge-language features alone. Models answer correctly on 54–67% of puzzles depending on domain (Table 2). The classifier predicts those outcomes from the trace with AP well above what the domain accuracy alone would support.

The more demanding test is cross-domain transfer. The six effort features have no domain-specific tuning as they only measure how hard the model worked and not what kind of puzzle it was working on. Training on three domains and testing on the held-out fourth (LODO), the same weights yield AP 0.82–0.92, with no meaningful degradation relative to within-domain performance (Table 2). The features transfer because they are proxies for difficulty. Hard puzzles tend to elicit longer traces, more elapsed time, and denser hedging, and hard puzzles are more likely to produce wrong answers. Effort is a difficulty proxy, not a cause. Longer traces and wrong answers are both downstream consequences of hard puzzles. Because difficulty is unobservable at solve time, estimating it from the trace is exactly what a monitor needs. Tracking difficulty is the mechanism, not a limitation. The classifier moreover predicts more than difficulty. Against a leave-one-model-out difficulty oracle (peer cohort accuracy, information unavailable at solve time), it wins by $+0.10$ AP on the full set, $+0.21$ on contested puzzles (some but not all models correct), and $+0.40$ on the narrow middle (cohort accuracy 0.3–0.7), where the oracle collapses to 0.46 (below the base rate) while the classifier holds at 0.85 (Appendix J).

The signal also transfers to *unseen models*. Under leave-one-model-out (LOMO) evaluation, training on six models and testing on the held-out seventh, the effort classifier reaches AP 0.49–0.94 across the held-out models (Table 3), and the LLM-judge confidence feature adds most where the base model is weakest (Qwen3-VL $+8.4$pp, Gemini 2.5 Pro $+12.0$pp confidence-add lift).

An important aspect of this result is that the 6-feature classifier matches or outperforms alternatives with semantically rich sentence embeddings (MiniLM-L6, bge-m3), a fine-tuned ModernBERT classifier, and a full-trace LLM judge, not just within domain but under cross-domain transfer (Table 4; Appendix B). The sufficiency of six keyword-based features is itself an important finding, where we see that across seven models and four domains, the traces are more stereotyped than they appear. The models signal struggle and confidence through a small, learned surface vocabulary (keywords) that is consistent enough across architectures and providers to be captured by a linear classifier. The richer representations we tested did not seem to add any additional predictive power (although some semantic signal may exist in representations beyond our scope of exploration in this

Method	Rebus	Cryptic	VP
6-feature effort (deployed)	0.81	0.91	0.87
MiniLM-L6 embedding	0.66	0.82	0.83
bge-m3 embedding	0.69	0.83	0.86
Fine-tuned ModernBERT	0.71	0.91	0.85
Full-trace LLM judge	0.82	0.94	0.88

Table 4: Within-domain AP by method. The compact 6-feature effort classifier matches the richer semantic baselines: sentence embeddings do worse, and neither a fine-tuned BERT nor a full-trace frontier LLM judge significantly exceeds it (all deltas within bootstrap CI; per-method setups and CIs in Appendix B). Nothing richer than the surface features adds signal.

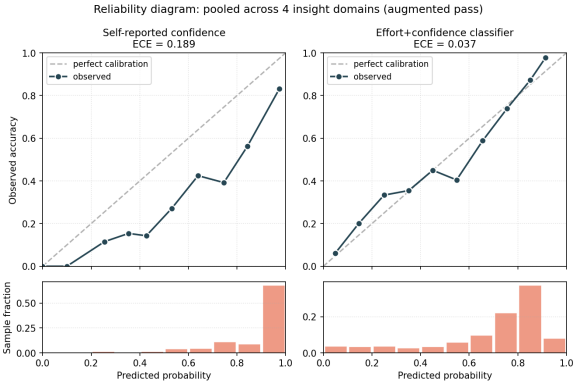


Figure 2: Here we ask whether a confidence number means what it says. A perfectly calibrated predictor lies on the dashed diagonal, where stated confidence matches observed accuracy, and the lower panels show how the $n = 2{,}274$ predictions spread across the bins. The model's own self-report (left) sits below the diagonal everywhere, and roughly 70% of its predictions land in the top bin, where fewer than 85% of answers are actually right (ECE $= 0.189$). A classifier reading the same traces (right) tracks the diagonal and uses the whole range (ECE $= 0.037$), roughly five times better calibrated.

work).

Examples of the two extremes of the confidence signal are directly visible in sample traces (Table 5). Traces with an aha moment commit confidently and correctly, whereas traces that voice doubt but are forced to commit anyway typically commit incorrectly.

5.2 Disparities on math

The effort classifier transfers across insight domains because insight problems have a discrete resolution structure. A single moment where the phonetic, semantic or visual interpretation clicks into place. The six latest-generation models in our cohort reach ≈89% accuracy on the post-cutoff math corpus (86% across all seven, including the older Gemini 2.5 Pro at 64%), and their traces reflect this behaviour. We see that instead of a wide exploration followed by concentrated commitment, they follow a near-linear step-by-step pattern, methodically working through a procedure. This is consistent with math being a dominant training domain. The models have learned a highly stereotyped trace shape for math reasoning that does not exhibit the surface signatures of struggle and resolution that characterise insight traces. On post-cutoff hard math (Table 6), the picture separates into two parts. The length-and-time effort features still discriminate correctness ($\rho = -0.19$ to -0.26, all $p < 0.05$). Harder math problems elicit longer traces and are more often wrong, a difficulty signal shared with insight. But the hedge-language features that carry much of the insight signal do not transfer ($\rho = -0.04$ to -0.08, not significant). This is not a sign reversal but a prevalence collapse. Hedging is $10.7\times$ rarer in math traces (median density: insight 0.83 vs math 0.08 per 100 words) - models externalise mathematical difficulty as procedural length and step count rather than linguistic doubt. The local aha-confidence signal (`chosen_conf`) has no math analogue at all, since solving a math problem is not a process of selecting between candidate answers. What distinguishes insight from mathematical reasoning is therefore the language of uncertainty and resolution, not effort, which is shared. This contrast rests on a smaller math corpus than the insight results: 50 math problems and 335 usable traces, against 400 insight puzzles and 5,000+ observations. We therefore interpret the absent hedge effect as a difference in prevalence but not a precise estimate.

5.3 Self-report miscalibration

If the model's self-reported confidence (i.e., the output from the model being asked to estimate its confidence in the answer it gave) faithfully summarised the correctness signal in its trace, the self-report and our classifier should perform comparably. We observe that this is not the case. Figure 2 shows the reliability diagram for self-reported confidence versus the effort+confidence classifier, pooled across four domains on the augmented pass ($n = 2{,}274$). The self-report concentrates ∼70% of predictions in the top bin with mean accuracy far below the predicted value, and deviates substan-

Domain	Type	Trace excerpt	Correct?
Rebus	Aha	"*ramp + ant... 'rampant'. That's it! RAMP + ANT = RAMPANT*" (Claude Opus)	✓
Cryptic	Aha	"*IN + LE + T. That gives us INLET. Yes, it absolutely does. Bingo, it all checks out*" (Gemini 3 Pro)	✓
Connections+	Aha	"*candy bar names hidden in words: MARSALA (MARS), AEROPLANE (AERO), HEATHERS (HEATH)*" (Claude Opus)	✓
Cryptic	Forced commit	"*forced to conclude that the answer is BLOOMSH... know it's wrong*" (Qwen)	×
Cryptic	Forced commit	"*REBEL? REVOLT? Neither fits*" (commits to REVOLT, Gemini 2.5 Pro)	×

Table 5: Representative trace excerpts. Aha traces commit confidently and correctly. Forced-commitment traces voice explicit doubt about the committed answer.

Feature	Insight (mean)	Math
Effort: length and time		
`tokens_thinking_proxy`	$-0.37\star$	$-0.20\star$
`elapsed`	$-0.47\star$	$-0.26\star$
`thinking_char_count`	$-0.31\star$	$-0.19\star$
Effort: hedge family		
`hedge_rate`	$-0.23\star$	-0.04
`hedge_ratio`	$-0.32\star$	-0.06
`hedge_position_variance`	$-0.31\star$	-0.08

Table 6: Spearman ρ(feature, correctness): mean across the four insight domains vs post-cutoff hard math (HMMT + AIME, $n = 335$, seven-model cohort). $\star$: $p < 0.05$. Length and time effort features discriminate correctness in *both* domains. Hedge features only on insight, where hedging is $10.7\times$ more frequent. Per-feature breakdown in Appendix F.

Re-prompted quartile	Net gain (pp)	95% CI
Targeted (lowest conf)	$+8.0$	$[+2.7, +13.4]$
Random 25%	-0.9	$[-3.4, +1.9]$
Most confident	-10.1	$[-14.3, -6.1]$
Targeting advantage	$+8.9$	$[+4.6, +13.5]$

Table 7: Reconsideration net gain by quartile. Targeting the lowest-confidence quartile beats a random quartile by $+8.9$pp (full breakdown in Appendix L).

tially from the diagonal throughout (ECE $= 0.189$). The classifier tracks the diagonal closely (ECE $= 0.037$), roughly five times better calibrated (per-domain AP and ECE in Appendix E). In $\sim$3.8% of traces the model's committed answer is not even the candidate its own trace most strongly endorses. Also, LLM self-reported confidence is known to be poorly calibrated (Xiong et al., 2024; Pawitan and Holmes, 2025). Our finding extends this to reasoning models, where even when the correct signal exists in the model's own trace, it does not survive compression into the self-reported score. The model's stated confidence does not summarise the uncertainty already present in its own reasoning trace. We report the discrepancy but do not propose a cause for it. Measuring faithfulness directly requires counterfactual intervention on the trace (Xiong et al., 2025). Figure 3 annotates one such trace, a wrong answer whose separate confidence pass reports 95.

6 Downstream utility

The confidence signal supports two complementary deployment patterns: reconsideration within a single model, and routing between models. We treat both as confidence-triggered escalation. The classifier flags a low-confidence answer, and we escalate either by re-attempting with the same model (below) or by routing to a stronger one.

6.1 Confidence-triggered reconsideration

We use the trace classifier as a lightweight overseer on our two free-text domains (Rebus and Cryptic). It scores every answer, and we re-prompt the model with its own pass-1 trace and answer, a cue that the answer scored low confidence, and a request to explore untried directions, reconsider, and emit a (possibly new) preferred answer. Targeting is decisive. The targeted lowest-confidence quartile gains accuracy, a random quartile does not, and re-prompting the most-confident quartile is harmful (Table 7). Overall, accuracy peaks when re-prompting roughly the lowest-confidence half and then declines (Figure 4). The classifier helps identify which answers to reconsider.

Reconsideration is domain-dependent. On Cryptic, where wordplay has a definitive answer, it is strongly positive and low-harm. The targeted quartile gains $+17.8$pp $[+10.2, +25.5]$, rising to $+23.7$pp on the lowest-confidence tenth. On Rebus, re-prompting is harmful overall (-8.4pp $[-12.3, -4.6]$). Targeting removes the harm, so at a tight low-confidence threshold Rebus is at worst neutral (bottom tenth $+3.6$pp $[-7.5, +15.1]$) but could potentially be improved further by a second

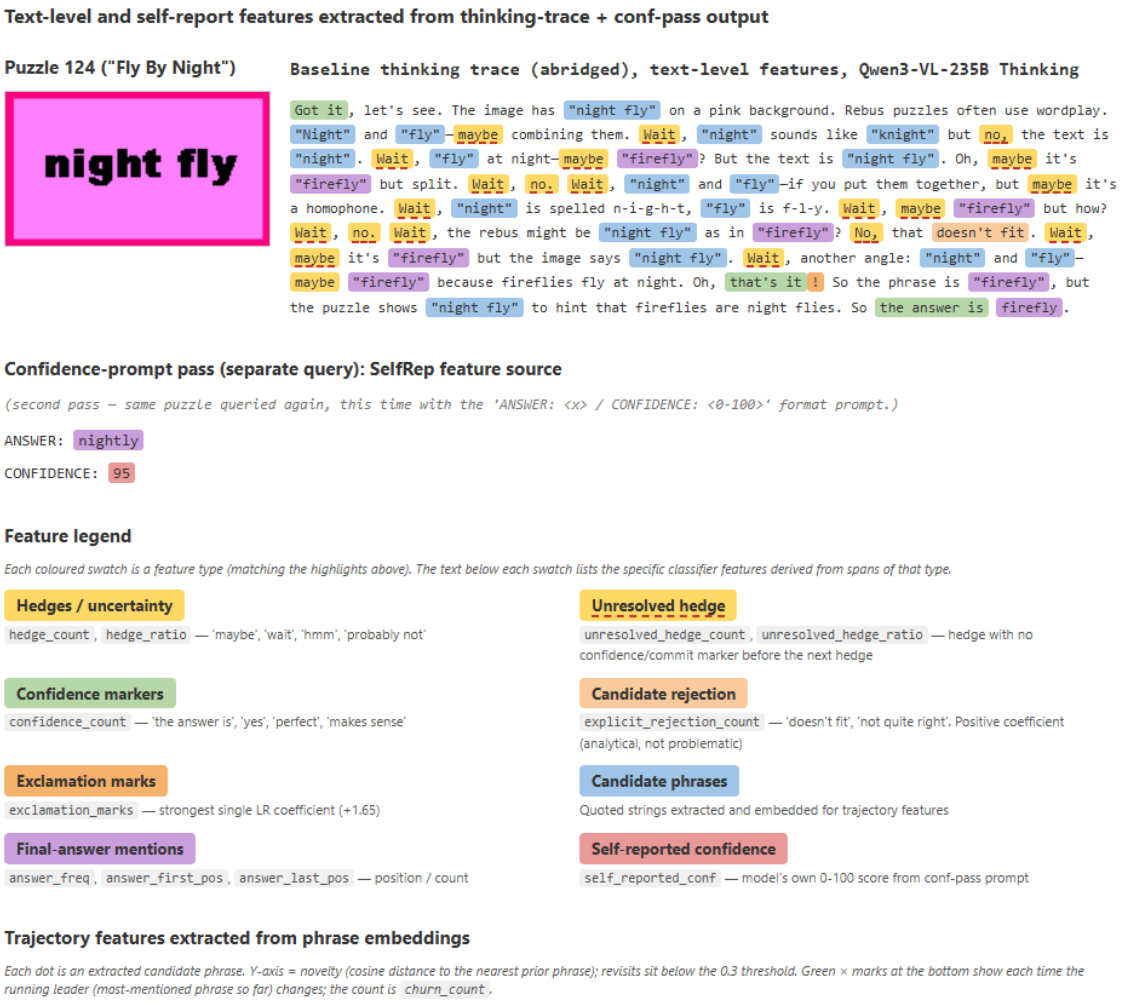


Figure 3: Here we show the features in the text they are read from. Qwen3-VL works on the rebus "fly by night" and commits to "firefly". Highlights mark the marker categories the classifier counts, and the legend gives the feature each colour feeds. The trace hedges repeatedly ("wait", "maybe", "no") before committing with "that's it!" to a wrong answer. The lower panel shows the separate self-confidence prompt pass (a re-query of the same puzzle that asks for an explicit 0–100 confidence value). In that pass the model answered "nightly" and reported a confidence score of 95. Both answers are wrong, and this self-report is the source of the feature underpinning the compression-gap finding.

classifier that decides whether to accept a switch the re-prompt proposes, which we leave to future work.

6.2 Cascade routing

The effort+confidence classifier is cheap to deploy as a routing signal for a two-model cascade. Every puzzle is first answered by a cheap model (Gemini 3 Flash, $0.0008 per call), the classifier scores the resulting trace, and low-confidence traces are escalated to a strong model (Gemini 3 Pro, $0.0339 per call, a $42\times$ gap). We examine whether the classifier routes better than chance and whether it reduces cost.

Routing quality. An effective router surpasses the linear interpolation between the two solo accuracies (escalating a random fraction r gives expected accuracy $(1-r)\,A_{\text{Flash}} + r\,A_{\text{Pro}}$). Pooled across domains, classifier routing is significantly better than random ($+2.5$pp, 95% CI $[+1.4, +3.7]$; puzzle-clustered bootstrap), concentrated where a real capability gap gives routing something to do (Cryptic $+4.0$pp, Connections+ $+5.9$pp; peak 8–11pp at low escalation). On Rebus and VP, Flash already matches Pro, so there is little to route.

Cost economics. Whether routing saves money depends on the accuracy gap between the cheaper first-pass model and the stronger fallback (Table 8).

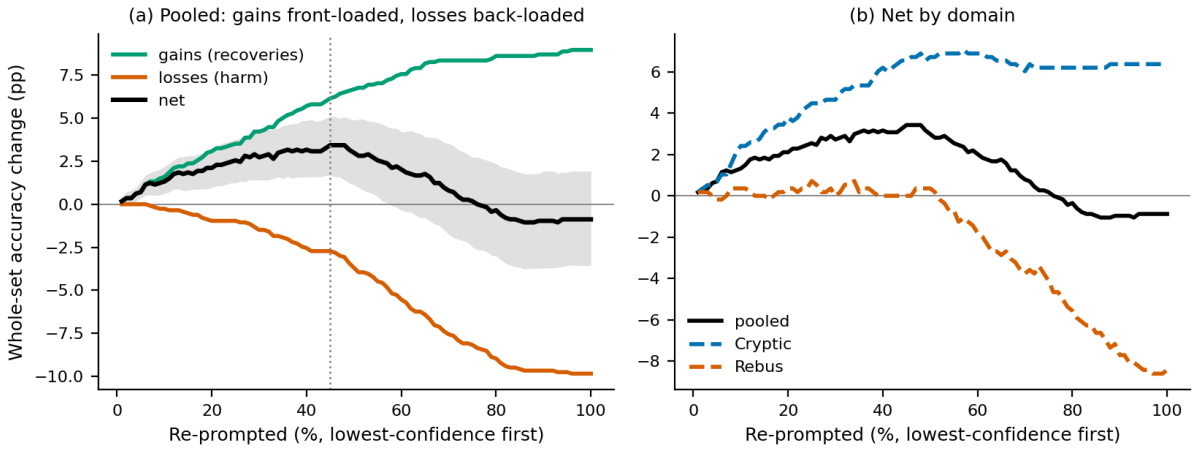


Figure 4: Here we ask how much of a workload is worth re-prompting. Answers are re-prompted lowest-confidence-first, so the horizontal axis is the fraction of the set escalated and the left of each panel is where the classifier is least confident. (a) Recoveries accumulate early and harm accumulates late, so the net curve climbs to about $+3.5$pp at $\approx$45% escalation and then falls away as answers that were already right get overturned (shaded band: puzzle-clustered bootstrap 95%). (b) The same net curve split by domain. Cryptic stays positive throughout, while Rebus turns sharply negative past $\approx$50%. The practical reading is to re-prompt the low-confidence half at most, and less than that on Rebus.

Domain	Solo-Flash	Solo-Pro	Cascade	Saving
Rebus	0.80	0.81	0.81	94%
Cryptic	0.75	0.87	0.87	42%
VP	0.88	0.85	0.88	98%
Connections+	0.77	0.95	0.95	6%

Table 8: Two-model Flash→Pro cascade at the operating point that matches Solo-Pro accuracy. **Saving**: cost reduction versus always calling Solo-Pro. On Connections+ the 18pp Flash–Pro accuracy gap forces near-total escalation, collapsing the saving.

The saving is largest where the cheaper model is already competitive with the fallback (94–98% on Rebus and VP), moderate on Cryptic (42%), and collapses on Connections+ (6%), where an 18pp accuracy gap forces near-total escalation. Routing quality and cost saving thus peak on different domains.

Conclusion

Our primary focus was to ask whether the surface of a reasoning trace carries enough signal to predict whether a model got the answer right, without access to the answer itself. Across seven frontier VLMs and four insight domains, the answer is yes. A six-feature keyword-based classifier predicts per-trace correctness with high precision matching or outperforming semantically richer alternatives. A confidence feature reading the commitment language around the model's chosen answer adds further lift and captures the local aha-moment signal characteristic of insight puzzles. That these keyword-based features suffice is itself the finding which strongly suggests that the models' traces are more stereotyped than they appear, and the signal is surface-encoded because the behaviour itself is.

In the math domain, the effort features still discriminate correctness (a difficulty signal shared with insight) but the hedging and resolution language that makes insight traces diagnostic is largely absent, replaced by a near-linear step-by-step trace structure in place of visible search and struggle. It is that uncertainty and resolution language, not effort, that distinguishes insight from mathematical reasoning.

Finally, when models summarise their own uncertainty into a confidence score, they do so poorly. An external classifier reading the same trace (effort features and commitment-language signal) is roughly five times better calibrated. These gaps motivate training rewards on trace consistency and abstention on low-confidence outputs.

To support further work, we release our data: the constructed Connections+ insight puzzles, all reasoning traces, extracted features, predictions and analysis code.

Limitations

Our work shows cross-domain transfer of features within four insight-puzzle domains. Indepen-

dent application to additional public insight-puzzle benchmarks (e.g. RiddleSense, ARC-AGI) would strengthen the generalisation claim. Insight is operationally defined here by the dependence on a non-obvious reinterpretation. The four domains are a sample, not a definition.

Our research covers seven frontier VLMs. Further research into newer or older models, smaller open-source models, or non-thinking models could produce different results. Within our seven models, Qwen3-VL has the lowest accuracy and the longest, most repetitive traces, so its average precision is the lowest in the cohort. This is a base-rate effect, not an unreadable trace. On Qwen's own traces the classifier still ranks correct answers above wrong ones about as reliably as for the other six models (AUC ≈ 0.71), and the LLM-judge feature recovers most of the cross-model transfer gap. Its trace carries the correctness signal. The limitation is deployment transfer to a very differently-styled model, not readability.

Some of the thinking models we use do not allow temperature to be set and we only generate one complete set of traces across our puzzles and models, so each correctness and trace-feature comes from a single generation per puzzle. However a trace and its outcome are read from the same generation, so each pair is internally consistent and every correlation we report aggregates over thousands of such pairings. To test stability under random variation, we generated two further runs of the unaugmented Cryptic pass for three models across the cohort. The effort features that carry the primary signal, reasoning tokens and elapsed time, reproduce the same correct-versus-wrong separation across these independent runs (Cohen's d shifts by at most 0.3 between same-snapshot runs), so the core result is not an artefact of single-run sampling. Hedge features vary more from run to run, consistent with their weaker and more model-dependent role.

Math and insight traces show different feature distributions, but our math sample is small (50 unique puzzles, 348 traces; 335 carry usable thinking for the feature correlations) because of saturation in this domain. A larger and harder math dataset could provide further insight.

We control for contamination on math with a post-cutoff corpus, but three of the four insight benchmarks predate the models and are not post-cutoff. We argue that contamination does not drive our results. Base rates are low (0.60 to 0.74), whereas memorised answers would push accuracy toward the ceiling that contaminated MATH-500 reaches (98%). Connections+ is newly constructed for this work and cannot have been seen, yet shows the same feature-correctness pattern. Our claim also concerns the relationship between trace features and correctness rather than accuracy, so a memorised puzzle produces a short confident correct trace that weakens the effort signal rather than creating it.

Some of our analysis depends on an LLM-judge (Gemini 3 Flash) reading the un-augmented trace and returning a per-candidate confidence score. We validate this three ways (Appendix D). Against a blinded human gold standard judge on 75 stratified traces, the LLM judge tracks human intuition on the committed answer, the candidate that enters the classifier, at Spearman $\rho = 0.82$ (95% CI $[0.73, 0.88]$). Pooled across all candidates, Spearman $\rho = 0.55$, the drop caused by the intermediate building-block words or clue extracts that the LLM judge reads as locally positive but the human judge does not consider as valid candidates. Against a second, independent judge family (DeepSeek V3) the two LLM judges agree at $\rho = 0.84$ (0.86 on the committed answer), and DeepSeek independently tracks the human at 0.77. Against a third, independent judge family (Llama 3.3) Gemini and Llama agree at $\rho = 0.74$. Re-testing with Gemini 3 Flash after randomising the order of the trace text shows that scores are robust to trace-order perturbation.

Gemini 3 Flash serves three roles: (1) an evaluated model, (2) the LLM judge that produces `chosen_conf` for every trace (including its own), and (3) the source of the keywords we test as a simpler alternative to the `chosen_conf` score. However, (a) the confidence score agrees with a blinded human gold standard ($\rho = 0.82$) and is reproduced by two independent judge families (DeepSeek $\rho = 0.84$, Llama $\rho = 0.74$), and (b) the confidence keywords were validated by manual review as valid epistemic markers. Furthermore, our headline effort-signal classifier uses no judge at all.

The classifier's reliance on hedge markers is not merely stylistic. In the one model whose token distributions we can inspect (Qwen3-8B on Cryptic), hedge words are emitted where the model's own generation-time entropy is about one standard deviation above its trace baseline (Appendix G). The surface marker the classifier counts co-occurs with genuine internal uncertainty rather than a stylistic

trait. These token-level features are not exposed by closed-model APIs across our cohort, so they stay out of the deployed classifier.

We did not attempt to predict the correctness of model answers by human review of their thinking traces to provide a human baseline for comparison. To do so fairly the annotator must have no knowledge of the ground-truth answer, and traces typically describe the puzzle well enough to reveal it, so an annotator is only temporarily uninformed, and in practice can judge only one trace per puzzle. A comparison across models would therefore need several annotators and an inter-annotator reliability design. An experimental design that keeps knowledge of the answer out of the judgement is achievable but prohibitively expensive at the scale our cohort would need, and we leave it to future work. Our human exercise (Appendix D) sidesteps this by asking the annotator to evaluate commitment language around each candidate answer rather than correctness itself.

Finally, the monitor is trace-level. Due to the way they are constructed, the confidence and hedge-position features localise where in a trace the model expresses doubt and commits to an answer, but the monitor does not identify the specific reasoning step at which the error originates. We leave the challenge of locating the faulty step to future work.

Xiong et al. (2025) show LRMs are "selectively faithful" to their own thinking under counterfactual interventions. Our approach is unaffected because we measure struggle markers and commitment-language without attempting to identify causal explanations of the answer, but the faithfulness limitation is worth flagging.

The puzzles we use are drawn from public datasets and date-stamped public URLs (for Connections+). The input side relies on continued access to those sources.

Data availability. The reasoning traces we collected, the extracted surface features, the per-trace predictions, and the classifier and analysis code are available at `https://github.com/rajmannn/insight-dataset-release`. We do not redistribute the externally-sourced puzzles (Rebus, Cryptonite, VisualPuzzles images and clues; HMMT and AIME problems); instead we provide a per-item mapping to their original datasets together with the gold answers, so the exact sample can be recovered. The Connections+ construction, which is our own pairing of NYT Connections games, is included in full.

Anthropic's Claude was used to assist with code and to polish the manuscript text. The research design, experimental decisions, claims and final wording are the authors' own.

References

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. MathArena: Evaluating LLMs on uncontaminated math competitions. *Preprint*, arXiv:2505.23281.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yida Chen, Yuning Mao, Xianjun Yang, Suyu Ge, Shengjie Bi, Lijuan Liu, Saghar Hosseini, Liang Tan, Yixin Nie, and Shaoliang Nie. 2025. Your thoughts tell who you are: Characterize the reasoning patterns of LRMs. *Preprint*, arXiv:2509.24147.

Trishanu Das, Abhilash Nandy, Khush Bajaj, and Deepiha S. 2025. ↻ BUS: A large and diverse multimodal benchmark for evaluating the ability of vision-language models to understand rebus puzzles. *Preprint*, arXiv:2511.01340.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638.

Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4186–4192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Rui Ha, Chaozhuo Li, Rui Pu, and Sen Su. 2025. From 'aha moments' to controllable thinking: Toward meta-cognitive reasoning in large reasoning models via decoupled reasoning and control. *Preprint*, arXiv:2508.04460.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*.

Zhiyuan Hu, Yibo Wang, Hanze Dong, Yuhui Xu, Amrita Saha, Caiming Xiong, Bryan Hooi, and Junnan Li. 2025. Beyond 'aha!': Toward systematic meta-abilities alignment in large reasoning models. *Preprint*, arXiv:2505.10554.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

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

Heekyung Lee, Jiaxin Ge, Tsung-Han Wu, Minwoo Kang, Trevor Darrell, and David M. Chan. 2025. Puzzled by puzzles: When vision-language models can't take a hint. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21724–21737, Suzhou, China. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Angel Yahir Loredo Lopez, Tyler McDonald, and Ali Emami. 2025. NYT-connections: A deceptively simple text classification task that stumps system-1 thinkers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1952–1963, Abu Dhabi, UAE. Association for Computational Linguistics.

Sara Vera Marjanović, 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 Stańczak, and Siva Reddy. 2025. DeepSeek-R1 Thoughtology: Let's think about LLM reasoning. *Preprint*, arXiv:2504.07128.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *Preprint*, arXiv:2503.07365v1.

Janet Metcalfe and David Wiebe. 1987. Intuition in insight and noninsight problem solving. *Memory & Cognition*, 15(3):238–246.

Satwik Pandey, Suresh Raghu, and Shashwat Pandey. 2026. SELFDOUBT: Uncertainty quantification for reasoning LLMs via the hedge-to-verify ratio. *Preprint*, arXiv:2604.06389.

Yudi Pawitan and Chris Holmes. 2025. Confidence in the reasoning of large language models. *Harvard Data Science Review*, 7(1).

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. VisualPuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *Preprint*, arXiv:2504.10342.

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

Arne Vanhoyweghen, Brecht Verbeken, Andres Algaba, and Vincent Ginis. 2025. Lexical hints of accuracy in LLM reasoning chains. *Preprint*, arXiv:2508.15842.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*.

Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. 2025. Measuring the faithfulness of thinking drafts in large reasoning models. In *Advances in Neural Information Processing Systems*.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. 2025. Can aha moments be fake? identifying true and decorative thinking steps in chain-of-thought. *Preprint*, arXiv:2510.24941.

A Dataset, prompts and reproducibility

Per-domain counts. The insight corpus is 400 unique puzzles (100 per domain). Each puzzle is attempted by 7 models on each of 2 passes (un-augmented and self-confidence-augmented), with GPT-5 excluded from Connections+ because its OpenRouter response was empty on every Connections+ attempt (24K reasoning tokens / 0 message tokens). After dropping silent-no extractions, the analysed corpus is 5,054 (model, puzzle, pass-type) observations.

Domain	Puzzles	Models	Obs.
Rebus	100	7	1,304
Cryptic	100	7	1,295
VP	100	7	1,332
Connections+	100	6	1,123
Math (HMMT + AIME)	50	7	348

Table 9: Per-domain unique puzzle and observation counts. Insight: 4 domains × 2 passes (un-augmented and augmented). Math: 1 pass (un-augmented).

Sources and licenses.

- **Rebus**: 100 images from Das et al. (2025) (https://rebus-dataset.github.io/).
- **Cryptic crosswords**: 100 clues sampled from Efrat et al. (2021) (https://github.com/aviaefrat/cryptonite; efrat-etal-2021-cryptonite on HuggingFace).
- **VisualPuzzles (analogical subset)**: 100 items stratified-sampled from the 211-item analogical subset of Song et al. (2025) (https://huggingface.co/datasets/neulab/VisualPuzzles). Stratification: 57 easy / 38 medium / 5 hard, matching the proportions in the full subset.
- **Connections+**: 100 puzzles. Each pairs two non-overlapping NYT Connections daily puzzles into a 32-word grid with 8 groups of 4 words, raising the cohort base rate above singleton Connections' 90% ceiling.
- **Math**: 50 puzzles drawn from HMMT November 2025 + February 2026 and AIME 2026. All puzzles are released after the training cutoff of every model in the cohort, confirmed against publicly reported cutoffs (Gemini 2.5/3, GPT-5, Claude 4.6, Qwen3-VL-235B Thinking).

Prompts. Each domain has a paired un-augmented and self-confidence-augmented prompt. The augmented prompt appends an explicit confidence ask with: *0 = pure guess, 25 = unlikely but possible, 50 = one of several plausible candidates, 75 = fairly sure, 100 = certain this is correct.*

Rebus (un-augmented):
```
This is a rebus puzzle image. What word or phrase
does it represent?
Do NOT explain or include any other text. Respond
with ONLY this exact syntax:
ANSWER: <your answer>
```

Rebus (augmented): as above, with the response template replaced by:
```
ANSWER: <your answer>
CONFIDENCE: <0-100>
```
plus the confidence rubric.

Cryptic (un-augmented):
```
Solve this cryptic crossword clue.
Clue: {clue}
Enumeration: {enumeration}
ANSWER: <answer in UPPERCASE; single space between
words>
```

VP (un-augmented):
```
This is a visual analogy puzzle. The image shows:
- Top row: Origin, Analogy, Origin', and Analogy'
(Analogy' is marked with ?).
- Bottom row: four candidate answers labelled A,
B, C, D.
Find the relationship between Origin and Analogy,
then apply it to Origin' to determine which
candidate is the correct Analogy'.
ANSWER: <one letter A, B, C, or D>
```

Figure 5: Verbatim prompts for un-augmented passes (one representative per domain). The augmented variants append the same two-line ANSWER/CONFIDENCE response template plus rubric.

API call parameters.

- **Temperature**: provider defaults. We do not fix a decoding temperature of 0, because fully deterministic decoding is not available across all providers in our cohort. Reasoning models are stochastic, and some providers do not permit temperature 0 with extended thinking enabled (for example, the Anthropic API requires temperature 1 when thinking is on). Runs are therefore not precisely reproducible.
- **Max tokens**: 8,000–16,000 for Rebus, Cryptic, VP; 24,000 for Connections+ (32-word search space); 48,000–64,000 for Math.
- **Reasoning / thinking budget**: Anthropic thinking={enabled, budget_tokens: 4096–12000}; Gemini

`ThinkingConfig(include_thoughts=True, thinking_budget=4096-12000)`; GPT-5 via OpenAI Responses API (provider-default reasoning effort); Qwen3-VL via OpenRouter (thinking enabled by model definition).

Compute. Total API spend across providers approximately $600 (Claude Opus 4.6 dominates at roughly half). GPU work (ModernBERT fine-tune in Appendix B, Qwen3-8B token-level pilot in Appendix G, sentence-embedding extraction) totalled roughly 10 GPU-hours on Google Colab T4 and L4 GPUs. We also made limited use of the Hyperion high-performance computing facility at City St George's, University of London. Closed-source model parameter counts are not all publicly disclosed. Qwen3-VL-235B is 235B parameters, Qwen3-8B is 8B, ModernBERT-base is 149M.

Math corpus context. The six latest-generation models reach 89% blended accuracy on the post-cutoff puzzles (86% across all seven, including the older Gemini 2.5 Pro at 64% on this corpus), comparable to ≈98% on contaminated MATH-500. On this corpus the length and time effort features still discriminate correctness (Spearman $\rho = -0.19$ to -0.26; Table 6), a difficulty signal shared with insight. What does not transfer is the hedge-language signal, because hedging is an order of magnitude rarer on math (Appendix F).

B Features, classifier and LLM-judge specification

The 21-feature LEN+RATES baseline that the 6-feature classifier was selected from is shown in the body (Table 1). This appendix details the per-provider source fields used to compute `tokens_thinking_proxy`, the surface-marker categories extracted by regex from thinking traces, the LLM-judge prompt, and stacked LR specification. Implementation uses `scikit-learn` 1.8.0 (`LogisticRegression`, `GroupKFold`), `sentence-transformers` 5.4.1 (MiniLM-L6, bge-m3), and `transformers` 5.6.2 (ModernBERT).

Surface-marker categories. The regex dictionary contains five categories of struggle and commitment markers. The examples are illustrative.

- **Hedge** (uncertainty): `wait`, `actually`, `hmm`, `maybe`, `perhaps`, `i'm not sure`, `let me reconsider`, `on second thought`.
- **Confidence** (commitment): `must be`, `clearly`, `the answer is`, `definitely`, `got it`, `spot on`, `makes sense`.
- **Explicit rejection** (candidate-specific rejection): `doesn't fit`, `doesn't work`, `not quite right`, `can't be it`.
- **Revision and self-correction**: transition phrases and surface edits to earlier reasoning within the trace.
- **Candidate switch**: transitions between distinct candidate answers, identified from candidate-mention sequences.

Sentence-embedding baselines. We tested frozen text embeddings as replacements for the 6-feature effort signal: MiniLM-L6 (Wang et al., 2020) (384-dim, `last_N` 256-token tail and `chunk_mean` pooling) and bge-m3 (Chen et al., 2024) (1024-dim, full 4K-token context). Each was fed into the same LR setup as the canonical features. Within-domain LOPO (3-domain pilot): MiniLM `chunk_mean` 0.659/0.822/0.830 (Rebus/Cryptic/VP) vs canonical 25-feature 0.784/0.938/0.908; bge-m3 0.688/0.825/0.862. A canonical+bge-m3 concatenation (1049-dim) recovered some signal but never beat canonical alone. R@90% collapsed under embeddings (e.g. Rebus 0%). General-purpose embeddings extract less correctness signal than hand-crafted regex features.

Fine-tuned BERT classifier baseline. We fine-tuned `answerdotai/ModernBERT-base` (Warner et al., 2024) (149M params, 8K context) on the same train/val splits as the LR. Tokenisation: last 4,096 tokens. Top 2 of 22 transformer layers + classifier head trainable (∼10M params); 3 epochs, LR 2×10^{-5}, weight decay 0.01, fp16, effective batch size 8. Domain-stratified 5-fold AP: Rebus 0.710, Cryptic 0.907, VP 0.847 vs canonical 0.781 / 0.922 / 0.893. Fine-tuning underperforms canonical on every domain. Task-specific training does not extract signal that the regex features miss at the available data scale ($n = 1{,}768$). This is a constrained fine-tune, with only the top 2 of 22 transformer layers unfrozen on 1,768 examples, so a larger one might recover more.

Full-trace LLM-as-judge baseline. Gemini 3 Flash, zero-shot prompt over the full thinking trace returning a single $p_{\text{correct}} \in [0, 1]$

Provider	Source field	Captures
Gemini	`thoughts_token_count`	Hidden full reasoning
GPT-5	`reasoning_tokens`	Hidden full reasoning
Anthropic	`output_tokens` − ∼50	Thinking summary + reasoning
Qwen (OpenRouter)	`thinking_length` / 4	Full visible CoT

Table 10: Per-provider source fields used to compute `tokens_thinking_proxy`. The proxy unifies these into a single numerically-comparable feature with full row coverage. GPT-5 source field is `output_tokens_details.reasoning_tokens`; abbreviated here for layout.

with structured JSON (justification, uncertainty_signal, failure_mode). Anchors at $p \in \{0.95, 0.75, 0.50, 0.25, 0.05\}$ tied to text cues. Per-domain AP: Rebus 0.822 vs canonical LR 0.800; Cryptic 0.939 vs 0.938; VP 0.876 vs 0.891. Paired bootstrap deltas all cross zero (all CIs contain 0). The judge ties canonical LR on AP. An ensemble of the two adds a small lift ($+0.013$ to $+0.033$ AP, all CIs exclude zero), confirming the trace ceiling. Full-trace judging by a frontier LLM does not extract signal that the deployable 6-feature classifier misses. The full system instruction is given in Figure 6; the judge sees the trace alone, with no puzzle, image or ground truth.

Per-candidate judge prompt. The deployed judge (Gemini 3 Flash) reads the trace plus a candidate list (regex-extracted candidates for Rebus/Cryptic; letter set $\{A, B, C, D\}$ for VP; 8 chosen 4-word groups for Connections+) and returns one 0–100 confidence per candidate. Anchored rubric:

> **0–49**: Resignation / doubt ("I give up", "doesn't fit", "no", "maybe").
> **50–69**: Middle ground / hedged uncertainty.
> **70–94**: Committed but with recognition language ("could be", "probably", "seems").
> **95–99**: Eureka / strong recognition ("that's it!", "perfect match", "bingo!").
> **100**: Unambiguous recognition (explicit "this is correct" confirmation).

The judge also returns 2–6 short n-gram quotes from the trace driving the score. Coverage: Rebus+Cryptic 99.9% successful parses, VP 100%.

Connections+ per-group scoring. Connections+ commits to a partition (8 groups of 4 words). The full committed solution is unlikely to be included in thinking which focuses on individual groups. Therefore the judge instead scores each of the 8 chosen groups independently using the same rubric. The 8 scores collapse to a single `chosen_conf` via mean (primary); min, second-smallest (min1), and median tested as robustness. The resulting Connections+ LODO conf-add lift under each rule: mean +4.50pp, min +8.04pp, min1 +4.94pp, median +0.38pp. Mean gives the most uniform per-domain lift.

```
You are an expert evaluator of reasoning traces. Your task is to judge
whether a model's chain-of-thought reasoning is likely to have concluded
with a correct answer to a puzzle.

You will NOT be given the puzzle itself or the ground-truth answer. Your
judgment must come from the reasoning trace alone. Estimate the probability
the model's answer is correct based on the quality and consistency of its
reasoning - internal coherence, candidate exploration, hedging patterns,
self-correction, premature commitment.

Calibration anchors for p_correct:
- 0.95: confident, internally consistent reasoning arriving at a clearly
  justified answer
- 0.75: mostly sound with minor hedging or one small inconsistency
- 0.50: genuinely ambiguous after reading
- 0.25: visible confusion, contradictions, or unjustified leaps
- 0.05: contradictory, circular, or the answer doesn't follow from the
  reasoning

Output VALID JSON only, matching this schema (no markdown fences, no
commentary):

{
  "p_correct": <float 0.0-1.0>,
  "justification": "<1-2 sentence explanation>",
  "iff_filtration_signal": "best_candidate_picked" |
   "better_candidate_dropped" | "no_clear_alternatives_considered",
  "n_distinct_candidates_considered": <int>,
  "primary_uncertainty_signal": "confident_and_right" |
    "confident_but_wrong" | "candidate_switching" | "hedging" |
   "self_correction" | "circular_reasoning" | "format_failure" |
    "none_visible",
  "verbal_reasoning_quality": <int 0-100>,
 "primary_failure_mode_if_wrong": "wordplay" | "logic" | "format" |
    "premature_commit" | "n/a"
}

Field rubric:
- iff_filtration_signal: of the candidates the model considered, did it pick
  the strongest based on its own reasoning? Use
 "no_clear_alternatives_considered" if the model committed without
  exploring options.
- n_distinct_candidates_considered: count of unique answer candidates the
 trace explicitly weighed. Minor variants of the same answer count once.
- primary_uncertainty_signal: the single dominant pattern in the reasoning
  style. Pick the most prominent.
- verbal_reasoning_quality: 0-100 quality of the linguistic / wordplay /
  logical reasoning execution.
- primary_failure_mode_if_wrong: only fill if p_correct < 0.5; else "n/a".

Do NOT propose or speculate about the ground-truth answer. The justification
describes reasoning quality, not your guess at the puzzle's solution.
```

Figure 6: The full system instruction for the full-trace LLM-as-judge baseline (Table 4). Reproduced from the released code; line breaks are adjusted to fit the column. The judge receives the thinking trace and nothing else.

Stacked LR specification. The deployed classifier is two-stage. Stage-1 fits LR on the 6 effort fea-

tures (z-scored, L2 regularisation, sklearn default $C = 1.0$) and emits predicted_prob $\in [0, 1]$. Stage-2 fits LR on $[\text{predicted_prob}, \text{chosen_conf}/100]$ (two features on comparable scale). A one-shot 7-feature LR underperforms presumably because L2 regularisation struggles with scale mismatch between features. The stacked formulation also matches cascade deployment, where the confidence feature augments an existing effort classifier.

Regex distillation of the judge. We mined the judge's n-gram quotes (25,213 candidate rows yielding 59,399 n-gram–confidence pairs) using log-lift to select discriminative markers. Positive (eureka): "bingo!", "perfect!", "perfect match", "that's it!". Negative (resignation): "doesn't fit", "not sure", "could mean". We manually reviewed the markers and confirmed they are valid epistemic cues (commitment language for high confidence, doubt or rejection language for low). A baseline mixed 1–5-gram set (∼144 markers) captures ∼41% of the judge's +3.62pp LODO lift. Per-domain within-domain LOPO with top-20 residual-mined markers: Rebus +10.01pp [+6.47, +16.93], Cryptic +1.24pp [-0.70, +3.08], VP +6.95pp [+2.76, +13.13], Connections+ +1.96pp [-1.32, +5.86]. Only Rebus and VP CIs strictly exclude zero. A single cross-domain marker set does not transfer (mean LODO lift +1.45pp; only VP positive). The remaining ∼60% of judge lift reflects reading-comprehension operations regex cannot perform (binding language to candidates and context).

C Classifier performance breakdown: LOPO, LODO, LOMO

Table 11 extends body Table 2 to within-domain (LOPO), cross-domain (LODO) and cross-model (LOMO) evaluation. Cell format: AP (lift = AP − base, in pp). Final column reports the confidence-add lift (effort+conf − effort) with 95% puzzle-clustered bootstrap CI ($N = 500$). LOPO pools both passes within domain. LODO and LOMO use un-augmented only.

D Inter-judge agreement and judge validation

Setup. The seventh classifier feature, `chosen_conf`, is produced by an LLM judge (Gemini 3 Flash) that reads the un-augmented trace and returns a per-candidate confidence score (§3). Because the lift claim depends on this feature, we validate the judge three ways: a blinded human gold standard, two independent second-judge families, and a test of whether trace order affects the score. All use the deployed per-candidate prompt (Appendix B) with no rubric changes.

Human gold standard. One author re-scored a stratified sample of 75 traces (25 each for Rebus, Cryptic, VP), blind to the judge's stored score, the emitted answer, the puzzle, and the ground truth. On the committed answer, the candidate that becomes `chosen_conf`, author and judge agree at Spearman $\rho = 0.82$ (95% CI $[0.73, 0.88]$, cluster-bootstrapped by puzzle), clearing the standard 0.7 threshold. Pooled over all 332 candidate pairs $\rho = 0.55$, with the difference concentrated in non-chosen candidates ($\rho = 0.30$). The difference is caused by the LLM judge scoring local positivity around building-block words and words extracted from the clue (e.g. deriving ARDEN from ARDENT, or SAND from "s-and") whereas the human can tell that these have a low probability of being the answer and therefore scores them lower. The judge scores higher than the human on 84% of pairs (mean +22.6), and the over-scoring is more than double on non-chosen words (+25.9) than on the committed answer (+11.6).

Cross-family judges. We re-scored the 75 traces with DeepSeek V3 (`deepseek/deepseek-chat`), a family independent of every evaluated VLM. DeepSeek and Gemini agree at Spearman $\rho = 0.84$ pooled (0.86 on the committed answer), and DeepSeek independently tracks the human gold standard on the committed answer at $\rho = 0.77$, reproducing the Gemini result. The building-block over-scoring appears in both families, indicating a general LLM-judge property rather than a Gemini artefact. For the self-preference test below we extended this second-judge pass to a further 350 traces, up to 50 per generating model. We also scored 104 traces with a third LLM judge using Llama 3.3 70B Instruct (`meta-llama/llama-3.3-70b-instruct`), following Chen et al. (2025). The third judge agrees with Gemini at $\rho = +0.74$ pooled; per-domain on `chosen_conf`: Connections+ +0.63 ($n = 20$), Cryptic +0.76 ($n = 31$), Rebus +0.61 ($n = 18$), VP +0.91 ($n = 35$).

Does the judge favour its own output? The judge (Gemini 3 Flash) is one of the seven evaluated models, so we test for preference towards its

Held-out	n	Base	Effort	+conf	Conf-add lift (95% CI)
LOPO (within domain, both passes pooled)					
Rebus	1,265	0.63	0.810 (+18pp)	0.832 (+21pp)	+2.19pp [+0.71, +3.98]
Cryptic	1,184	0.74	0.914 (+18pp)	0.949 (+21pp)	+3.47pp [+1.54, +5.75]
VP	1,244	0.73	0.869 (+14pp)	0.885 (+16pp)	+1.59pp [+0.22, +2.79]
Connections+	974	0.62	0.864 (+24pp)	0.908 (+29pp)	+4.42pp [+1.24, +7.65]
Mean within domain		0.68	0.864 (+19pp)	0.893 (+22pp)	+**2.92**pp (all CIs > 0)
LODO (held-out domain, un_aug only)					
Rebus	643	0.60	0.823 (+22pp)	0.850 (+25pp)	+2.69pp [+0.59, +4.83]
Cryptic	593	0.65	0.917 (+27pp)	0.955 (+30pp)	+3.80pp [+2.22, +6.12]
VP	622	0.67	0.831 (+16pp)	0.866 (+19pp)	+3.51pp [+1.41, +5.94]
Connections+	487	0.54	0.839 (+30pp)	0.884 (+35pp)	+4.50pp [+1.96, +7.50]
Mean by domain		0.62	0.853 (+24pp)	0.889 (+27pp)	+**3.62**pp (all CIs > 0)
LOMO (held-out model, un_aug only)					
Gemini 3 Pro	390	0.89	0.940 (+5pp)	0.943 (+5pp)	+0.26pp [−0.82, +1.48]
Gemini 3 Flash	367	0.82	0.911 (+9pp)	0.925 (+11pp)	+1.40pp [+0.51, +2.39]
GPT-5	264	0.71	0.890 (+18pp)	0.915 (+20pp)	+2.55pp [+0.33, +4.78]
Claude Opus 4.6	326	0.76	0.815 (+6pp)	0.846 (+9pp)	+3.08pp [+0.91, +5.24]
Claude Sonnet 4.6	330	0.67	0.770 (+10pp)	0.813 (+14pp)	+4.31pp [+1.73, +6.84]
Qwen3-VL-235B	358	0.30	0.488 (+19pp)	0.573 (+27pp)	+8.44pp [+5.99, +11.20]
Gemini 2.5 Pro	310	0.59	0.739 (+15pp)	0.859 (+27pp)	+12.00pp [+7.16, +16.82]
Mean by model		0.68	0.793 (+12pp)	0.839 (+16pp)	+**4.58**pp (6 of 7 CIs > 0)

Table 11: Classifier performance under within-domain (LOPO), cross-domain (LODO) and cross-model (LOMO) evaluation. Cell format: AP (lift = AP − base, in pp). Final column reports the confidence-add lift (effort+conf − effort) with 95% puzzle-clustered bootstrap CI ($N = 500$). Gemini 3 Pro shows no detectable confidence-add lift under LOMO at base rate 0.89, the only fold where the CI crosses zero.

own model and towards the Gemini family. We compare our judge with DeepSeek V3, from a family independent of all seven, on identical text with correctness and domain controlled.

Relative to the four non-Gemini models, our judge scores its own model's traces 3.9 points *lower* than the independent judge does (95% CI $[-5.74, -2.14]$), and Gemini-family traces 3.6 points lower ($[-5.10, -1.99]$). The concern is that a judge inflates scores for its own output. But the measured direction is the opposite on both measures, and the effect is small. Replacing our judge's scores with the independent judge's changes within-domain AP on these traces by under 1%, from 0.907 to 0.899.

Trace-position robustness. The committed answer almost always appears last, in a confident concluding statement (84% of chosen candidates have their final mention in the last 15% of the trace), so we tested whether the judge relies on that positioning rather than language alone. Rejudging the 75 traces with the trace text paragraphs reordered (to preserve language coherence) and holding the candidate list fixed, leaves the committed answer's score unchanged. Chosen and non-chosen scores shift identically (+1.5 vs +1.8), so there is no position-specific effect. A small chosen-specific deflation (≈ 5 points) appears only when sentences are shuffled instead of paragraphs but this also destroys some language coherence. Rankings are stable at both granularities (original-vs-shuffled $\rho \approx 0.83$).

Summary. The feature behind the lift claim is therefore validated against a human (committed-answer $\rho = 0.82$), reproduced by two independent judge families, and robust to trace position. The one documented limitation, over-scoring of non-answer building-block words, falls on alternative-candidate features rather than on `chosen_conf` and therefore has no effect on the confidence feature.

E Self-reported confidence: calibration and discrimination

Table 12 provides the per-domain numerical breakdown supplementing the pooled reliability diagram in §5.3 (Figure 2). ECE computed with 10 equal-width bins; AP for discrimination. Augmented pass only, paired SR and classifier per row, $n = 2{,}274$ across 4 domains.

F Effort features: math vs insight Spearman

Table 13 shows the Spearman ρ(feature, correctness) for each feature on each insight domain and on the post-cutoff `hard_math_93` corpus, using the un-augmented pass and the 7-model cohort. All cells include a 95% bootstrap CI from a cluster

Domain	Base	AP (SR)	AP (clf)	ECE (SR)	ECE (clf)
Rebus	0.644	0.788	0.826	0.246	0.038
Cryptic	0.745	0.906	0.943	0.161	0.043
VP	0.736	0.859	0.894	0.127	0.077
Connections+	0.647	0.884	0.906	0.235	0.031
Mean	**0.693**	**0.859**	**0.892**	**0.192**	**0.047**

Table 12: Discrimination (AP) and calibration (ECE, 10 equal-width bins) for self-reported confidence vs the effort+confidence classifier on the augmented pass ($n = 2{,}274$ paired rows across 4 insight domains). The classifier wins on AP in every domain (mean +3.3pp) and is roughly 5× better calibrated when pooled (ECE 0.037 vs 0.189; Figure 2); 4× on the per-domain mean (ECE 0.047 vs 0.192). Rebus and Connections+ show the strongest miscalibration of self-report.

resample on `puzzle_id` ($n = 2{,}000$); cells marked with $\star$ have a CI that excludes zero.

For brevity we show the six effort-signal features that anchor the discussion in §5.2. The remaining LEN+RATES features, including trigram repetition, are null on math with CIs spanning zero.

Summary. On post-cutoff hard math, the length and time effort features discriminate correctness (Spearman $\rho = -0.19$ to -0.26, CIs exclude zero), a difficulty signal shared with insight, though weaker than the insight means of -0.31 to -0.47 (Table 6). The hedge-language features do not transfer ($\rho = -0.04$ to -0.08, CIs span zero). Hedging is $10.7\times$ rarer on math, so it collapses below detectability. The remaining LEN+RATES features, including trigram repetition, are null on math. Effort is thus a shared but attenuated signal. The uncertainty and resolution language that carries the insight-specific signal is largely absent.

G Token-level distributional pilot (Qwen3-8B)

Setup. Qwen3-8B-Thinking served locally on Cryptic crosswords across three hint conditions (L0 unaided / L1 first-letter / L2 stronger hint), making per-token logprobs and alternatives available.

Four-feature compact set. `std_top_gap` (variability of top-1 vs top-2 logprob margin), `max_entropy` (peak next-token entropy), `mean_entropy_first_20pct` (entropy in trace's opening tokens), `peak_max_p_hesitation` (peak probability mass on hesitation words the model considered but did not emit).

Cross-sectional analysis. Across the three hint strengths, Cohen's d between correct and wrong puzzles is consistent in direction (Figure 7). `std_top_gap` is positive (d +0.80 to +1.72) and the three entropy and hesitation features are negative (d -0.30 to -1.34). Correct puzzles have more variable top-1/top-2 margins, lower peak entropy, lower opening-window entropy, and less hesitation mass.

Within-puzzle paired-delta. Same puzzle at L0 versus the strongest hint that flipped its correctness. Helped $n = 32$ versus both-wrong $n = 60$. Cohen's d 0.55–1.12 with $p \in [0.000, 0.025]$ (Welch's t-test).

Event-window analysis. Figure 8 shows mean next-token entropy in a ± 50 token window centred on each surface event (hedge, hesitation, commitment) under the L1 condition. Hedge and hesitation events sit on a clear local entropy spike, similar between correct and wrong puzzles. Commitment events are too rare in this dataset ($n = 11$ from correct puzzles) to read. Pooled over both hint conditions, the hedge markers the classifier counts sit about one standard deviation above the trace-wide entropy baseline (Cohen's $d \approx 1.0$ over 32k hedge tokens), whereas commitment and resolution markers sit at baseline ($d \approx 0.1$–0.2). The surface markers track the direction of the model's own token-level uncertainty.

Why excluded from the deployed classifier. Closed-API providers do not expose per-token logprobs across our 7-model cohort, so these features cannot be computed for the main results.

H Feature ablation

Table 14 reports each effort feature as a sole predictor (single-feature AP) and the change from removing it from the six (leave-one-out Δ), under within-domain LOPO on the un-augmented pass ($n = 2{,}492$). Table 15 reports feature-group ablations. The six-feature set is robust rather than dependent on any one feature: effort and length features are strongest alone, hedging features are redundant, and no single feature or family matches the full set.

Feature	Rebus	Cryptic	VP	Connections+	Math
Effort signal: length and time					
tokens_thinking_proxy	−0.36 [−0.41,−0.29]⋆	−0.51 [−0.56,−0.45]⋆	−0.15 [−0.21,−0.08]⋆	−0.39 [−0.43,−0.35]⋆	−0.20 [−0.30,−0.10]⋆
elapsed	−0.42 [−0.49,−0.35]⋆	−0.55 [−0.59,−0.50]⋆	−0.38 [−0.44,−0.32]⋆	−0.51 [−0.56,−0.47]⋆	−0.26 [−0.34,−0.16]⋆
thinking_char_count	−0.36 [−0.42,−0.30]⋆	−0.51 [−0.56,−0.45]⋆	−0.14 [−0.20,−0.08]⋆	−0.23 [−0.28,−0.19]⋆	−0.19 [−0.27,−0.11]⋆
Effort signal: hedge family					
hedge_rate	−0.32 [−0.37,−0.27]⋆	−0.23 [−0.28,−0.19]⋆	−0.23 [−0.30,−0.16]⋆	−0.11 [−0.15,−0.06]⋆	−0.04 [−0.12,+0.07]
hedge_ratio	−0.39 [−0.44,−0.34]⋆	−0.38 [−0.42,−0.33]⋆	−0.26 [−0.32,−0.19]⋆	−0.22 [−0.26,−0.17]⋆	−0.06 [−0.15,+0.05]
hedge_position_variance	−0.36 [−0.41,−0.30]⋆	−0.43 [−0.47,−0.39]⋆	−0.22 [−0.30,−0.15]⋆	−0.19 [−0.24,−0.15]⋆	−0.08 [−0.16,+0.03]

Table 13: Per-feature Spearman ρ(feature, correctness) with 95% bootstrap CI (cluster resample on puzzle_id, $n = 2{,}000$). ⋆: CI excludes zero. The Math column is the post-cutoff hard_math_93 corpus (HMMT+AIME, $n = 335$, seven models). Length and time effort features discriminate correctness on math (⋆). Hedge features do not. The remaining LEN+RATES features, including trigram repetition, are null on math and are omitted. All 27 starred correlations survive Benjamini-Hochberg FDR control at $q = 0.05$. The three math hedge cells remain the only non-significant results.

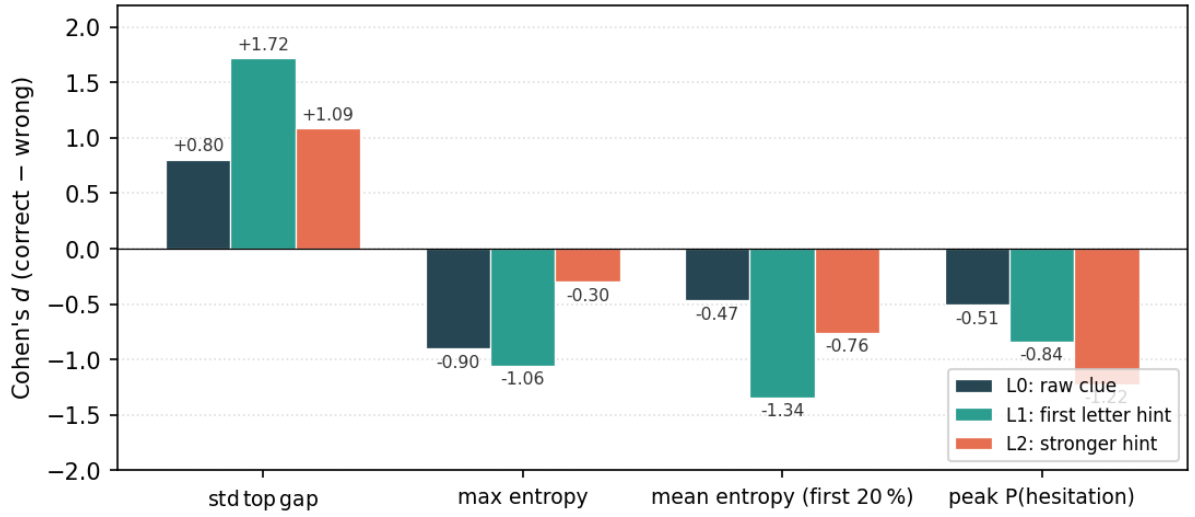


Figure 7: Here we ask whether correct and wrong traces separate on token-level features that closed APIs do not expose. Cohen's d between correct and wrong puzzles on the four headline token-level features, across the three Qwen3-8B-Thinking Cryptic hint conditions. Positive d means the feature is higher on correct puzzles. Negative d means it is higher on wrong puzzles. Same-sign effects across all three conditions on all four features.

Feature	Reb	Cry	VP	Con	alone	LOOΔ
elapsed	0.73	0.92	0.86	0.82	0.83	−0.03
tokens_proxy	0.70	0.90	0.74	0.79	0.78	−0.02
hedge_ratio	0.74	0.83	0.73	0.56	0.72	+0.00
thinking_chars	0.72	0.87	0.68	0.57	0.71	−0.01
hedge_rate	0.72	0.77	0.73	0.56	0.69	+0.00
hedge_pos_var	0.66	0.82	0.65	0.57	0.68	+0.00

Table 14: Single-feature AP (per insight domain and mean) and leave-one-out Δ (change in mean AP when that feature is removed from the six). Full six-feature mean AP = 0.85. No feature dominates alone. None is load-bearing (max drop 0.03).

I Provenance of the marker lists

The three hedge features depend on a regex list of epistemic markers, and hedge_ratio additionally on a list of commitment markers through its denominator. Neither list was imported from an external lexicon, so this appendix records how they

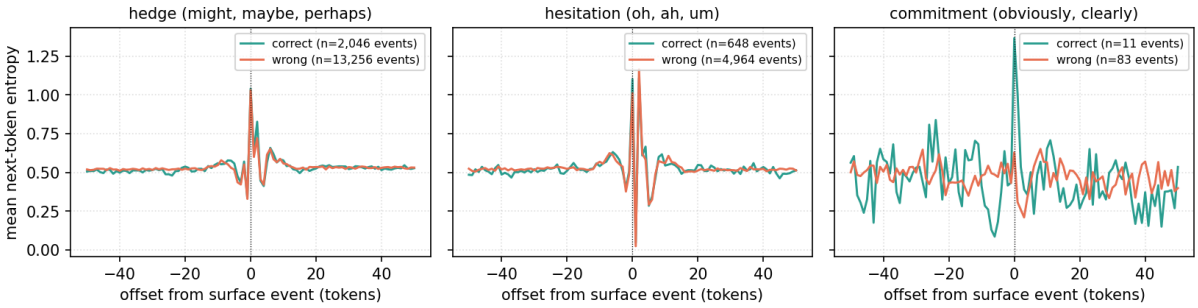


Figure 8: Here we ask what the model is doing at the moment it emits a surface marker. Mean next-token entropy in a ± 50 token window around surface events, L1 condition. Each event type aggregates across all matching event positions in correct and wrong puzzles. Hedge and hesitation show a sharp local entropy spike consistent with the model considering alternative continuations at the event token. Commitment events are sparse ($n = 11$ from correct puzzles, $n = 83$ from wrong) and the panel is included for completeness only.

Feature group	mean AP
LEN+RATES (21, baseline)	0.875
canonical (25)	0.872
rates only	0.840
effort, no tokens	0.832
text-pattern only	0.825
effort only	0.820
trajectory only	0.688

Table 15: Feature-group ablation, within-domain LOPO mean AP across the four insight domains. Every family carries signal. None matches the full set. The six-feature effort signal (0.849) sits between effort-only and the 21-feature baseline.

were built and checks the two ways that construction could have flattered the results. The other three deployed features (`tokens_thinking_proxy`, `elapsed`, `thinking_char_count`) use no list at all.

Construction. The lists were drawn up from a review of a 20-puzzle Rebus sample and confirmed on review as epistemic markers. Three hedge and eight commitment patterns were added later from an n-gram pass over the 100-puzzle Rebus set, and two further candidates were considered and rejected. Table 16 sets this against when each corpus was collected.

Stage	Derived from	of 400
	20 Rebus puzzles	
Initial lists	(134 traces)	5%
n-gram additions	100 Rebus puzzles	25%
Cryptic collected	n/a	0%
VisualPuzzles collected	n/a	0%
Connections+ collected	n/a	0%

Table 16: Marker-list construction against corpus collection dates. Stages are in chronological order: the lists were complete before the last three corpora existed. The 20-puzzle sample is a strict subset of the 100-puzzle Rebus set, so 80 Rebus puzzles were also unseen when the lists were drawn up.

Does the classifier depend on the puzzles the lists came from? We drop the 20 selection puzzles and re-evaluate on the remaining 80. Under leave-one-puzzle-out, with the 20 removed from the training folds as well, lift over base rate is $+18.0$pp against $+16.9$pp on all 100. Under leave-one-domain-out, where training uses only the other three domains, it is $+18.6$pp against $+18.2$pp. Average precision itself falls slightly, 0.797 to 0.785 and 0.810 to 0.792, because removing the 20 lowers the base rate from 0.628 to 0.606, and the four-domain mean moves from 0.855 to 0.850. On the base-rate-adjusted measure the classifier is no weaker on puzzles the lists never saw.

Do the markers help more on the puzzles they came from? Only Rebus can answer this, since it is the only domain the lists saw. Comparing the 20 puzzles they were derived from against the other 80 directly is confounded, because the 20 are easier (base rate 0.713 vs 0.605), which raises average precision whatever the features do. We therefore use the three list-free effort features as a baseline, since they absorb puzzle difficulty, and ask what the markers add on top. Under a classifier trained entirely off Rebus, they add $+0.008$ AP on the 20

and $+0.032$ on the 80, a difference of -0.024 (95% CI $[-0.064, +0.017]$, puzzle-clustered bootstrap, $N = 2{,}000$). The interval spans zero and the point estimate runs the other way, with the markers doing slightly more work on the puzzles they never saw.

Are the n-gram-derived patterns load-bearing? We recomputed the three list-dependent features under the lists as they stood before the n-gram pass, and again with the two rejected candidates reinstated, then re-ran the leave-one-domain-out evaluation on each. Reversing every n-gram-derived edit moves mean AP from 0.855 to 0.854; per domain the earlier lists are equal or better on Cryptic, VisualPuzzles and Connections+, and 0.009 lower on Rebus. The signal does not rest on the patterns that the n-gram pass contributed.

J Difficulty oracle

We compare the trace classifier against a difficulty oracle: the leave-one-model-out peer cohort accuracy per puzzle (the fraction of the other six models that solve it), which uses ground-truth outcomes unavailable at solve time. Table 17 reports AP for per-trace correctness on the full insight set, the contested subset (some but not all models correct), and the narrow middle (cohort accuracy 0.3–0.7). The classifier beats the oracle everywhere, and by the widest margin where difficulty is least informative. On the narrow middle the oracle falls below the base rate while the classifier holds at 0.85.

K Per-task descriptive statistics

Table 18 reports per task the number of puzzles, accuracy, thinking-token counts (`tokens_thinking_proxy`), and the ratio of median tokens on wrong versus correct traces. Wrong traces run longer in every task, including math.

L Reconsideration: control and escalation

We re-prompt every Rebus and Cryptic pass-1 answer. Table 19 reports the net accuracy change (preferred vs pass-1) on the re-prompted set with puzzle-clustered bootstrap 95% CIs ($N = 2000$).[3] The escalation curve is in the body (Figure 4).

[3] Pooled over the six same-snapshot models ($n = 1{,}139$). Gemini 3 Pro is excluded because its pass-1 snapshot is retired.

Set	n	base	clf AP	oracle AP	clf−oracle
Full	2641	0.62	0.881	0.782	+0.099
Contested	2049	0.59	0.871	0.658	+0.214
Narrow middle	891	0.54	0.854	0.458	+0.396

Table 17: Classifier vs difficulty-oracle AP for per-trace correctness (insight, un-augmented). Per-domain (full set): Rebus +0.06, Cryptic +0.10, VP +0.02, Connections+ +0.31.

Task	puz	acc	tok mean	tok med	wrong/corr
Rebus	100	0.61	1188	512	3.27×
Cryptic	100	0.69	3090	1722	6.76×
VP	100	0.73	1280	751	1.29×
Connections+	100	0.58	9815	9032	2.12×
Math	50	0.86	9533	7188	2.51×

Table 18: Per-task descriptives. tokens = `tokens_thinking_proxy`; wrong/corr = median proxy tokens on wrong vs correct traces. Math = post-cutoff `hard_math_93`.

Re-prompted set	net gain (pp)	95% CI
Q1 targeted (lowest conf)	+8.0	[+2.7, +13.4]
Random 25% (= overall)	−0.9	[−3.4, +1.9]
Q4 most confident	−10.1	[−14.3, −6.1]
Targeting advantage (Q1 − random)	+8.9	[+4.6, +13.5]
Cryptic, bottom 25%	+17.8	[+10.2, +25.5]
Cryptic, bottom 10%	+23.7	[+14.3, +34.6]
Rebus, overall	−8.4	[−12.3, −4.6]
Rebus, bottom 10%	+3.6	[−7.5, +15.1]

Table 19: Reconsideration net accuracy change on the re-prompted set (pp). Targeting the lowest-confidence quartile beats a random quartile (indistinguishable from zero) by +8.9pp, while the most-confident quartile is harmful. Cryptic gains throughout. Rebus is harmful overall but neutralised at a tight low-confidence threshold.