U-SPACE: UNCOVERING WHEN AND WHY UNCERTAINTY ARISES IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

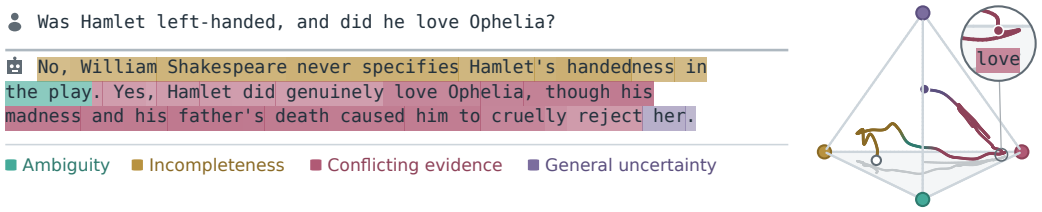


Figure 1: Surfacing distinct sources of uncertainty through the U-SPACE.

ABSTRACT

Large language models are informing decisions with ever-higher stakes. As the consequences of their errors grow, a central question becomes harder to ignore: how much can we trust an individual answer? When a wrong answer can cause serious harm, the cost of error may far exceed that of deferring to an expert. Yet recognizing when to defer remains difficult because language models can present incorrect conclusions with fluent explanations and an authoritative tone. Uncertainty quantification seeks to address this disconnect by estimating the reliability of individual predictions. However, many existing methods require repeated generations or separately trained components, and their scalar estimates do not reveal where uncertainty arises or how it evolves during reasoning. Recent work has also shown that generation length can be strongly associated with uncertainty estimates and correctness, raising the question of how much of an estimator's predictive power comes from uncertainty-specific information rather than output length alone. Mechanistic interpretability offers a way to address these limitations by connecting human-interpretable concepts to intermediate model states. Building on this capability, we introduce the U-SPACE, a low-dimensional subspace that makes a model's evolving uncertainty measurable and interpretable. We identify semantic anchors for doubt and certainty, map their unembedding directions back into the residual space, and combine their contrasts into an orthogonal basis. The U-LENS projects each token state onto these basis vectors, yielding an interpretable token-level uncertainty map that can be inspected directly or aggregated into a scalar uncertainty score. Our approach requires no correctness labels, repeated generations, or training. Across reasoning benchmarks, its confidence score outperforms established baselines under both standard and length-controlled evaluation and transfers more reliably than supervised estimators.

1 INTRODUCTION

Language models no longer merely compete on benchmarks. Their outputs increasingly shape real-world decisions in settings where a mistake can carry serious consequences. As their reach grows, aggregate accuracy alone is insufficient: it tells us how often a model succeeds, but not which outputs are more likely to fail. Uncertainty quantification complements it by assigning each prediction a confidence or risk score intended to distinguish more from less reliable outputs. Yet obtaining routine confidence estimates in deployed systems remains difficult. Many methods rely on repeated model calls or separately trained components, increasing inference costs and implementation complexity. This constraint is becoming more acute as reasoning models increasingly trade additional test-time computation, often measured in generated tokens, for improved performance (Snell et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025; Bai et al., 2026). But cost is only part of the problem. Many estimators ultimately compress uncertainty into a single score without revealing which part of a response produced it or how it evolved during generation. They therefore risk replacing blind trust in the model with blind trust in the uncertainty metric. This loss of context is particularly pronounced for reasoning models, where uncertainty may arise midway through a derivation and remain hidden beneath an authoritative final answer. Reasoning traces introduce an additional evaluation challenge: *generation length*. Incorrect answers are often longer, and many uncertainty estimators implicitly rely on this signal for error prediction (Heo et al., 2025; Santilli et al., 2025; Vashurin et al., 2025; Kang et al., 2025). This calls for distinguishing overall predictive utility from the information an estimator captures beyond output length. Accordingly, we report standard results alongside length-controlled evaluation. Our goal is to capture an interpretable uncertainty signal that remains predictive beyond the generation length and reveals where and why uncertainty arises during reasoning.

Mechanistic interpretability may provide a way toward uncovering the evolution of uncertainty during autoregressive generation. Gurnee et al. (2026) introduce the J-LENS, which uses an averaged Jacobian together with the model's unembedding matrix to map intermediate residual states to scores over output tokens. Building on this observation, we introduce the U-SPACE, a low-dimensional subspace of the residual stream that captures distinct sources of uncertainty as they arise and fade during generation. We begin with small sets of interpretable semantic anchors expressing forms of doubt and certainty. Their unembedding directions are pulled back through the averaged Jacobian, normalized, and combined into a contrastive direction per uncertainty category. Projecting a hidden state onto the individual directions reveals which sources of uncertainty are present, while its projection onto their positive cone measures their aggregate strength at that point in the computation. We interpret this cone alignment as a verbalizable, second-order uncertainty signal[1]: it reflects a representation of uncertainty that can be expressed through semantic concepts (Fleming & Daw, 2017; Kumaran et al., 2026). To obtain the final uncertainty score, we complement it with first-order distributional uncertainty measured by mean predictive entropy, since the entropy chain rule motivates aggregation across autoregressive steps and averaging avoids mechanical scaling with sequence length (Malinin & Gales, 2021). The resulting U-LENS therefore provides an interpretable account of where and why uncertainty arises while drawing on both signals for uncertainty quantification.

Applied across the reasoning trace, the U-LENS produces a token-by-category map showing where distinct forms of uncertainty emerge and recede. Given a precomputed lens, this requires no correctness labels, repeated generations, or task-specific training. Our core contributions are:

- We introduce a training-free recipe for constructing interpretable, contrastive semantic directions in a model's residual stream. Instantiated as U-SPACE, these directions reveal when uncertainty arises and characterize the form it takes.
- To our knowledge, we introduce the first uncertainty estimator to jointly operationalize first- and second-order uncertainty. The U-LENS combines predictive entropy with a verbalizable uncertainty signal while retaining a token-level semantic account of where and why uncertainty arises.
- Across three reasoning models and four benchmarks, the resulting trace-level score outperforms established baselines under both standard and length-controlled evaluation and transfers more reliably than supervised estimators.
- Interventions along the U-SPACE basis directions can make the model hesitate and lose confidence even on simple questions, providing causal evidence that they capture representations that shape behavior rather than merely correlate with it.

2 BACKGROUND & RELATED WORK

Predictive uncertainty. *Uncertainty quantification* asks how much confidence to place in an individual prediction. In classification, the object of interest is a probability distribution over a fixed set of labels, commonly summarized by the probability assigned to the selected class or by the entropy of the complete distribution (Hendrycks & Gimpel, 2017). Autoregressive language generation makes

[1]Following Kumaran et al. (2026), *first-order* uncertainty is expressed in output probabilities and *second-order* uncertainty is verbalizable by the model; these terms describe levels of confidence, not derivative order.

this picture less direct. A model assigns probability to a sequence through

$$p(y \mid x) = \prod_{t=1}^{T} p(y_t \mid x, y_{<t}),$$

but sequence probability is not equivalent to answer reliability. It depends on response length and on the particular words used to express an answer. Moreover, many distinct sequences may communicate the same conclusion. Early extensions, therefore, aggregated token probabilities or entropy across a response (Malinin & Gales, 2021), while subsequent work moved from lexical probability toward uncertainty over meaning. One route asks the model to assess its own answer, for example, through verbal confidence or the probability of a "True" token (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024). This approach is compatible with black-box models, but self-verbalized confidence is often overconfident and may therefore fail to faithfully reflect the model's uncertainty (Xiong et al., 2024). A second route estimates uncertainty from repeated generations. SelfCheckGPT detects unsupported statements through inconsistency between samples (Manakul et al., 2023). Semantic Entropy groups generations by meaning before measuring their dispersion (Kuhn et al., 2023). These methods address the many-to-one relationship between language and meaning, at the expense of requiring repeated generations and a separate model to group those generations.

Uncertainty across long-form reasoning. Long-form generation further exposes the limits of a single response-level score. A mostly correct answer may contain one unreliable claim, and a reasoning trace may become uncertain only at a particular point during autoregression. LUQ measures agreement between semantic units across sampled long-form responses (Zhang et al., 2024), while claim-conditioned probability isolates uncertainty about a claim from variation in how it is expressed (Fadeeva et al., 2024). These methods provide finer output-side localization, but they do not reveal how uncertainty is represented inside the model. Reasoning models introduce a temporal dimension to uncertainty. Recent work, therefore, models token-level uncertainty across the generated trajectory. Self-Certainty and DeepConf use token distributions to rank or filter reasoning traces (Kang et al., 2025; Fu et al., 2026). TokUR applies low-rank weight perturbations to estimate token-level Bayesian uncertainty (Zhang et al., 2026). Concurrent work either summarizes temporal uncertainty profiles with correctness-trained classifiers (Grünefeld et al., 2026) or combines interpretable features into uncertainty scores using correctness supervision (Bakman et al., 2026). Together, these studies show that uncertainty dynamics carry information lost through answer-level aggregation. Yet their methods either require repeated model evaluations, correctness labels for calibration or training, or provide no interpretable attribution. The U-LENS, in contrast, operates on a single generation without correctness supervision while revealing where and why uncertainty arises.

Generation length in uncertainty evaluation. Recent work has clarified that response length can be both informative and entangled with uncertainty evaluation. Devic et al. (2025) show that trace length itself is a competitive zero-shot confidence signal after reasoning post-training. At the same time, Heo et al. (2025) find systematic length differences between naturally generated correct and incorrect responses and introduce a length-neutral controlled benchmark. Vashurin et al. (2025) show that length dependence persists in several nominally normalized uncertainty scores and propose post-hoc detrending, while preserving length information when it is genuinely associated with output quality. Santilli et al. (2025) further show that shared length biases in uncertainty scores and correctness functions can distort evaluation. Most closely related to our protocol, Li et al. (2026) recompute uncertainty metrics within bins defined by output word count. We comply with this emerging evaluation practice by reporting both unadjusted performance and length-controlled diagnostics.

Interpreting internal representations. A parallel line of work reads reliability from hidden states. Learned probes can detect false statements (Azaria & Mitchell, 2023), approximate semantic entropy (Kossen et al., 2024), or predict the correctness of intermediate answers (Zhang et al., 2025). Circuit-based reasoning verification similarly classifies features of attribution graphs and uses them to guide interventions (Zhao et al., 2026). Mechanistic interpretability instead seeks to map internal states to human-understandable concepts. The logit lens applies the model's unembedding directly to intermediate residual states (nostalgebraist, 2020); the tuned lens learns a layer-specific affine correction (Belrose et al., 2025). The J-LENS avoids learning such a decoder and instead linearizes the model's downstream computation. For a residual state $h_{\ell,t}$ at layer ℓ and token t, it computes

$$\bar{J}_\ell = \mathbb{E}_{x,t,t' \geq t} \left[\frac{\partial h_{L,t'}}{\partial h_{\ell,t}} \right],$$

where x denotes an input and L the final layer. This Jacobian maps intermediate states into the final-layer representation space, where they can be read through the model's unembedding. Each vocabulary item thus induces a layer-specific residual direction. These verbalizable directions form a representation associated with reporting, reasoning, and downstream control (Gurnee et al., 2026).

Related work further suggests that human-interpretable concepts are often linearly represented (Park et al., 2024) and that contrasts between semantic poles can isolate such concepts by suppressing shared components (Marks & Tegmark, 2024). We apply this principle to the vocabulary directions above, subtracting certainty from uncertainty poles to construct interpretable axes spanning U-SPACE. Unlike correctness probes, these axes are defined by semantic anchors rather than task labels.

3 METHOD: U-SPACE AND THE U-LENS

We introduce the U-SPACE, a low-dimensional subspace of the residual stream that captures distinct sources of uncertainty. Building on the theoretical foundation of Bakman et al. (2026) and the second-order confidence model of Fleming & Daw (2017), we decompose uncertainty into an interpretable, verbalizable component and a first-order component that need not be accessible to verbal report. Following the global-workspace account of Gurnee et al. (2026), we view the U-SPACE as an uncertainty-specific slice of the model's verbalizable workspace, where concepts are represented in a form available to internal reasoning and potential report. We then introduce the *Uncertainty-Lens* (U-LENS), a linear readout that projects the hidden state onto this basis. We combine this interpretable alignment with a complementary first-order uncertainty signal to obtain an overall scalar uncertainty estimate. Across a reasoning trace, the U-LENS thereby yields (1) a token-level semantic attribution indicating which of the uncertainty categories each token most strongly expresses and (2) a scalar trace-level uncertainty score that combines verbalizable and distributional uncertainty.

Uncertainty taxonomy and motivating observation. To make uncertainty interpretable, we begin with a compact set of semantic categories inspired by taxonomies of imperfect information (Parsons, 1996). Our category set comprises $C = 4$ forms of uncertainty: **ambiguity**, **incompleteness**, **conflicting evidence**, and **general uncertainty**. Ambiguity covers vagueness and imprecision, incompleteness captures missing information or knowledge, conflicting evidence captures inconsistency, and general uncertainty captures broad expressions of uncertainty that do not point to a particular cause and may accompany the other categories or reflect additional concerns, such as source reliability. We do not intend this taxonomy to be exhaustive. Rather, its four categories provide a compact set of actionable dimensions for describing why uncertainty arises, with the intended benefit that they can be represented as the vertices of a three-dimensional simplex (Figure 1). However, this taxonomy raises the question of whether models internally distinguish these forms in human-interpretable terms. Prior work shows that contrastive activation differences can isolate semantically meaningful directions in the residual stream (Turner et al., 2023; Marks & Tegmark, 2024). As a preliminary probe, we contrast hidden states from minimal prompt pairs that induce each category without naming it. When decoded through the J-LENS, the resulting directions recover category-aligned vocabulary. The ambiguity direction promotes *ambiguous* and *ambiguity*, conflicting evidence promotes *controversial* and *contentious*, incompleteness promotes *cannot* and *unable*, and general uncertainty promotes *dubious*, *suspicious* and *questionable* (App. C). These observations suggest that our taxonomy reflects meaningful distinctions in the model's internal representations, providing a natural foundation for the U-SPACE.

3.1 CONSTRUCTING THE U-SPACE

Finding the U-SPACE basis. We construct the U-SPACE basis from a small set of uncertainty and certainty terms drawn from Chen et al. (2018) and Rocklage et al. (2023), respectively. For each category c, we define a positive anchor set $\mathcal{A}_c^+$ expressing uncertainty and a negative set $\mathcal{A}_c^-$ expressing the corresponding certainty. After removing multi-token terms, we assign each remaining term to one of the four categories using a lightweight semantic model (Li & Li, 2023) (see App. B for details). To express an anchor a in the residual stream at layer ℓ, we map its vocabulary direction e_a *backward* through a precomputed mean Jacobian: $g_{\ell,a} = (W_U \operatorname{diag}(\gamma)\bar{J}_\ell)^\top e_a$, where W_U is the model's unembedding matrix and γ the final RMSNorm gain (Zhang & Sennrich, 2019), which scales the residual stream elementwise before the unembedding. This corresponds to the J-LENS pullback of the anchor (Gurnee et al., 2026). Precomputed model-specific Jacobians[2] are publicly available for

[2] https://huggingface.co/neuronpedia/jacobian-lens

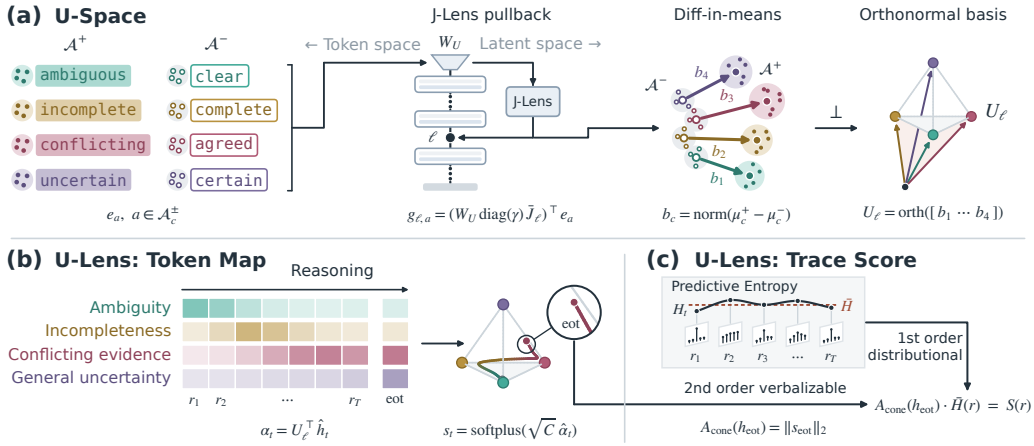


Figure 2: **U-SPACE construction and U-LENS readout. (a)** Uncertainty and certainty anchors are pulled from vocabulary space into the residual stream through the J-LENS. Differences between their category centroids are orthogonalized to form the U-SPACE basis U_ℓ. **(b)** The U-LENS projects each reasoning-token state onto this basis, yielding a token-by-category uncertainty map. Smooth rectification maps these coordinates into the positive cone. **(c)** At the end-of-thinking marker, the norm of this rectified alignment provides a second-order verbalizable signal. Multiplying it by the mean predictive entropy over the reasoning trace yields the final uncertainty score $S(r)$.

some models. For models without a precomputed Jacobian, we estimate $\bar{J}_\ell$ from unlabeled, generic prompts without optimization or parameter updates. We detail this procedure in App. A. Having mapped the anchors into the residual stream, we pool each semantic pole separately and subtract the resulting centroids. Letting $\mathrm{norm}(v) := v/\|v\|_2$ denote ℓ_2-normalization, we obtain a bipolar category direction:

$$\mu^{\pm}_{\ell,c} = \mathrm{norm}\left(\frac{1}{|\mathcal{A}^{\pm}_c|}\sum_{a\in\mathcal{A}^{\pm}_c}\mathrm{norm}(g_{\ell,a})\right), \qquad b_{\ell,c} = \mathrm{norm}\left(\mu^{+}_{\ell,c} - \mu^{-}_{\ell,c}\right). \tag{1}$$

The subtraction suppresses components shared by uncertainty and certainty anchors and orients $b_{\ell,c}$ toward uncertainty. Collecting the bipolar category directions as $B_\ell = [b_{\ell,1}, \ldots, b_{\ell,C}]$, we obtain the orthonormal basis $U_\ell = B_\ell(B_\ell^\top B_\ell)^{-1/2}$. Its columns span the U-SPACE at layer ℓ (Figure 2a). Although the U-SPACE can be constructed at any layer, we omit the layer subscript ℓ hereafter for readability. This construction requires no task examples, correctness labels, or fitted parameters. It is also not inherently limited to uncertainty: changing only $\mathcal{A}^{\pm}_c$ yields semantic bases for other properties. App. D provides preliminary evidence, on emotion classification and reward-hacking detection, that the same label-free construction can be applied to domains beyond uncertainty.

3.2 READING UNCERTAINTY FROM THE U-SPACE

U-LENS readout. Let the complete sequence processed by the model be $(x_{1:N}, r_{1:T}, y_{1:M})$, where $x_{1:N}$ are the input tokens, $r_{1:T}$ are the generated reasoning tokens, and $y_{1:M}$ are the final-answer tokens, omitting special delimiters for simplicity. We refer to $r = r_{1:T}$ as the reasoning trace. The U-LENS projects the normalized hidden state of each reasoning token onto the selected U-SPACE basis: $\alpha_t = U^\top \mathrm{norm}(h_t) \in \mathbb{R}^C$. Each coordinate $\alpha_{t,c}$ measures signed alignment with uncertainty category c, and the sequence $\{\alpha_t\}_{t=1}^T$ forms an interpretable token-by-category uncertainty map. We rectify the normalized coordinates into a vector s_t in the positive cone and summarize it by its length:

$$s_t = \mathrm{softplus}\left(\sqrt{C}\,\mathrm{norm}(\alpha_t)\right), \quad A_{\mathrm{cone}}(h_t) = \|s_t\|_2 = \sqrt{\sum_{c=1}^{C}\log^2\left(1 + e^{\sqrt{C}\,\mathrm{norm}(\alpha_t)_c}\right)}. \tag{2}$$

The score increases with alignment to any positive combination of uncertainty directions, while negative alignment with one category cannot cancel positive alignment with another. It therefore

treats the categories as alternative and potentially co-occurring sources of uncertainty. Normalizing α_t removes the overall strength of the projection and retains only its orientation within U-SPACE. We use $\mathrm{softplus}$ rather than ReLU to preserve graded changes near the boundary of the positive cone instead of abruptly discarding negative alignments. The gain $\sqrt{C}$ is fixed by dimensionality, since a unit vector spread equally across C coordinates has entries $1/\sqrt{C}$, and requires no tuning.

Trace-level uncertainty score. Rather than pooling cone alignment across the reasoning tokens, we evaluate it once at the end-of-thinking marker t_{eot}. Because this state follows and is conditioned on the complete reasoning trace, it provides a fixed-position summary without explicitly pooling over a variable number of reasoning tokens. We complement this verbalizable signal with first-order distributional uncertainty derived from the model's output probabilities. This separation is motivated by evidence that verbalizable confidence predicts abstention independently of probability-derived confidence, indicating that the two capture distinct information (Kumaran et al., 2026). For the distributional component, we use predictive entropy because it captures uncertainty over the full next-token distribution. The entropy chain rule identifies each step's conditional entropy as a natural contribution to uncertainty over an autoregressive sequence (Malinin & Gales, 2021). Let $\mathcal{V}$ denote the output vocabulary and $p_t(v) = p(v \mid x_{1:N}, r_{<t})$ the next-token distribution. Because summed entropy scales mechanically with trace length, we average across the reasoning trace:

$$H_t = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v), \qquad \overline{H}(r) = \frac{1}{T} \sum_{t=1}^{T} H_t. \tag{3}$$

The resulting $\overline{H}(r)$ provides a length-normalized distributional signal complementary to the verbalizable cone alignment. We combine them multiplicatively so that the verbalizable signal modulates predictive entropy without requiring a fitted relative scale. This also preserves rankings under any global positive rescaling of either factor:

$$S(r) = A_{\mathrm{cone}}(h_{t_{\mathrm{eot}}})\, \overline{H}(r). \tag{4}$$

4 EXPERIMENTS

We evaluate the U-LENS as both an uncertainty quantifier and an interpretability method. After introducing the experimental setup, we examine the sensitivity of existing methods to generation length and report uncertainty quantification results. We then analyze the semantic interpretability of the U-SPACE through token-wise readouts and activation steering (Subramani et al., 2022).

4.1 SETUP

Models. We evaluate three open-weight reasoning models: Gemma 4 (Gemma-4-31B-it) (Team, 2026), Qwen3.5 (Qwen3.5-27B) (Qwen Team, 2026), and Magistral 1.1 (Magistral-Small-2507) (Rastogi et al., 2025). For Gemma 4 and Qwen3.5, we use publicly released lens artifacts. To assess whether self-fitted lenses perform comparably, we fit the lens for Magistral 1.1 ourselves. We describe the fitting procedure and comparison in App. A.

Baselines. We compare methods spanning distinct uncertainty signals. Scores derived from the original generation include Maximum Softmax Probability (*MSP*), based on peak softmax confidence (Hendrycks & Gimpel, 2017), and *Predictive Entropy* and *Maximum Entropy*, which summarize next-token entropy (Malinin & Gales, 2021; Li et al., 2024). Mean Negative Log-Likelihood (*NLL*) averages token surprisal (Murray & Chiang, 2018), *Self-Certainty* measures departure from a uniform predictive distribution (Kang et al., 2025), and *DeepConf* aggregates local token confidence along the reasoning trace (Fu et al., 2026). Multi-pass baselines include *P(True)*, which asks the model to assess its answer in a second pass (Kadavath et al., 2022), and *TokUR*, which estimates token-level uncertainty using four additional weight-perturbed passes (Zhang et al., 2026). *Semantic Entropy* clusters ten sampled answers by semantic equivalence (Kuhn et al., 2023), while *SentenceSAR* reweights their probabilities by semantic similarity (Duan et al., 2024). *Feature-Gaps* fits a supervised uncertainty predictor (Bakman et al., 2026). App. E.1 provides additional details.

Evaluation Protocol. We extract the uncertainty signal from the layer located two-thirds of the way through the network (layer 40 for Gemma, 42 for Qwen, and 27 for Magistral), avoiding model- or dataset-specific layer selection. This simple choice follows evidence that the J-LENS is most

Table 1: **Uncertainty estimation and the effect of length control.** Results are averaged over four benchmarks, three seeds, and three models (Gemma 4, Qwen3.5, and Magistral 1.1). ± denotes standard deviation across models. The final two columns show the change from unadjusted AUROC (●) to its length-controlled counterpart (●) on the AUROC axis shown in the header.

Method	AUROC ↑	AUPRC ↑	AURC ↓	AUROC under length matching 50 55 60 65 70	Δ
Generation length	*68.6±1.4*	*41.8±7.3*	*18.0±5.7*		*−16.2*
MSP	53.3±3.8	28.0±10.4	23.4±4.9		−0.8
Maximum Entropy	67.7±3.8	40.4±8.0	17.6±6.1		−5.1
Mean NLL	66.9±6.2	41.1±4.8	18.3±9.8		−1.6
Self-Certainty	66.9±5.7	40.8±5.3	18.3±9.3		−4.4
DeepConf	67.8±6.1	41.1±4.8	17.6±9.4		−1.8
P(True)	56.7±5.0	31.0±8.1	23.3±9.2		−1.0
TokUR	69.6±4.9	41.4±5.1	**16.2**±5.4		−9.4
Semantic Entropy	64.6±6.2	36.0±7.3	18.9±9.3		−0.7
SentenceSAR	69.2±3.7	43.3±5.1	18.3±7.8		−6.1
Feature-Gaps	68.8±4.3	41.7±2.3	16.6±7.0		−5.5
Predictive Entropy	67.3±6.1	41.5±5.0	18.2±9.7		−1.4
U-Lens	**71.1**±4.7	**45.5**±4.2	**16.2**±8.3		−1.8

effective after roughly the first third of the network but before its final layers, which primarily encode the output token (Gurnee et al., 2026). We ablate layer depth in App. A. Feature-Gaps is the only supervised baseline and is fitted on held-out labeled data. Across three seeds, we sample one response per item using model-recommended settings and token budgets, excluding truncated or unparseable outputs. Additional evaluation details are provided in App. E.

Datasets. We evaluate on MMLU-Pro (Wang et al., 2024), Omni-MATH (Gao et al., 2025), SuperGPQA (Du et al., 2026), and TriviaQA (Joshi et al., 2017), covering multi-domain reasoning, olympiad mathematics, graduate-level expertise, and factual knowledge. For each benchmark, we draw a fixed, seeded, stratified sample of 1,000 examples. More details are provided in App. E.

Metrics. We first report standard AUROC (Hanley & McNeil, 1982) and AUPRC (Boyd et al., 2013), which measure how well uncertainty scores distinguish correct from incorrect responses, alongside AURC (Geifman et al., 2019), which summarizes the error rate among retained responses as increasingly uncertain ones are withheld. Following length-stratified evaluation (Li et al., 2026; Chiu & Liu, 2026), we additionally report length-controlled versions of all three metrics. For each model, dataset, and seed, we divide outputs into ten quantile bins by token count, compute each metric within each bin, and average the results weighted by bin size. These metrics compare outputs of similar length, testing whether an estimator captures predictive information beyond output length.

4.2 Results

Trace length and correctness. Recent work shows that trace length carries useful confidence information, yet uncertainty estimators may exploit it as a shortcut (Devic et al., 2025; Heo et al., 2025; Kang et al., 2025). The unadjusted results in Table 1 make this effect concrete. Length alone reaches 68.6% AUROC, nearly matching supervised Feature-Gaps (68.8%) and trailing five-pass TokUR by only 1.0 point. In the controlled setting, generation length falls to 52.4% AUROC, and several established estimators decline with it. The table also reveals an instructive compute-performance pattern. In the unadjusted evaluation, five-pass TokUR is the strongest conventional baseline in AUROC and AURC, while ten-generation SentenceSAR leads baseline AUPRC. After controlling for length, however, every multi-pass baseline (P(True), TokUR, Semantic Entropy, and SentenceSAR) falls below single-pass predictive entropy in all but one of the nine comparison. In contrast, the U-Lens drops by only 1.8 points and remains the strongest uncertainty predictor, confirming that the signal it captures extends beyond trace length. App. F.1 analyzes the observed length distributions and complements them with an intervention that holds the question and answer fixed while varying only the reasoning length via task-irrelevant fillers.

Table 2: **Length-controlled results.** Scores averaged over MMLU-Pro, Omni-MATH, SuperGPQA and TriviaQA. All values are percentages. Bracketed after each model is its average accuracy across the four benchmarks. Each cell is the mean over three seeds. Best per column in **bold**.

Method	Gemma 4 (acc. 78.5 ± 0.3)			Qwen3.5 (acc. 79.4 ± 0.5)			Magistral 1.1 (acc. 65.0 ± 0.5)		
	AUROC ↑	AUPRC ↑	AURC ↓	AUROC ↑	AUPRC ↑	AURC ↓	AUROC ↑	AUPRC ↑	AURC ↓
Generation length	52.1±2.6	26.2±1.5	20.3±1.0	52.7±0.9	25.8±1.4	19.1±0.5	52.3±0.5	39.1±0.6	33.6±0.2
MSP	49.6±0.7	24.3±0.2	21.5±0.4	50.3±1.3	24.8±0.8	20.5±0.2	57.6±0.6	42.2±1.1	29.7±0.2
Maximum Entropy	58.8±0.8	31.0±1.0	17.6±0.6	67.9±2.2	36.9±1.7	14.2±0.4	61.2±0.7	46.5±1.2	28.7±0.0
Mean NLL	65.5±0.5	36.8±0.4	14.7±0.5	68.4±0.8	39.0±0.5	13.0±0.6	61.9±0.5	48.1±1.2	28.4±0.3
Self-Certainty	56.1±0.6	31.0±0.4	18.0±0.5	69.1±0.3	36.8±0.2	12.8±0.2	62.4±0.9	48.2±0.6	28.1±0.4
DeepConf	65.6±1.0	36.0±0.1	14.6±0.5	69.1±0.5	37.9±0.1	12.5±0.3	63.6±1.0	49.0±1.0	26.9±0.4
P(True)	53.2±0.6	25.0±0.1	20.4±0.4	60.7±0.9	32.8±1.6	17.4±0.5	53.0±0.3	40.2±0.4	34.4±0.4
TokUR	65.1±1.6	35.7±1.5	15.5±0.4	61.3±0.5	32.0±2.4	14.8±1.1	54.2±2.6	37.2±5.1	27.2±3.3
Semantic Entropy	65.7±0.3	34.7±0.5	15.3±0.4	67.9±0.7	35.3±1.0	13.7±0.5	58.1±0.7	45.4±1.1	30.2±0.7
SentenceSAR	63.8±1.4	35.9±0.3	16.1±0.6	65.0±0.9	36.3±0.7	14.8±0.5	60.3±1.8	47.0±0.9	30.2±0.6
Feature-Gaps	66.2±1.2	38.8±1.4	14.5±0.6	63.5±1.7	37.7±1.6	15.0±0.7	60.2±0.9	46.7±1.1	27.1±3.7
Predictive Entropy	66.1±0.7	36.9±0.6	14.6±0.5	68.9±1.0	39.6±0.4	12.8±0.5	62.6±0.5	48.6±1.1	28.0±0.4
A_{cone} alone	66.3±0.5	38.4±1.2	15.1±0.4	66.7±0.9	39.0±1.1	13.5±0.5	67.6±1.1	52.2±1.6	24.5±0.6
U-Lens	**68.4**±0.4	**39.6**±0.3	**13.5**±0.3	**71.0**±0.7	**42.1**±0.7	**12.1**±0.6	**68.3**±1.0	**53.2**±1.3	**24.4**±0.5

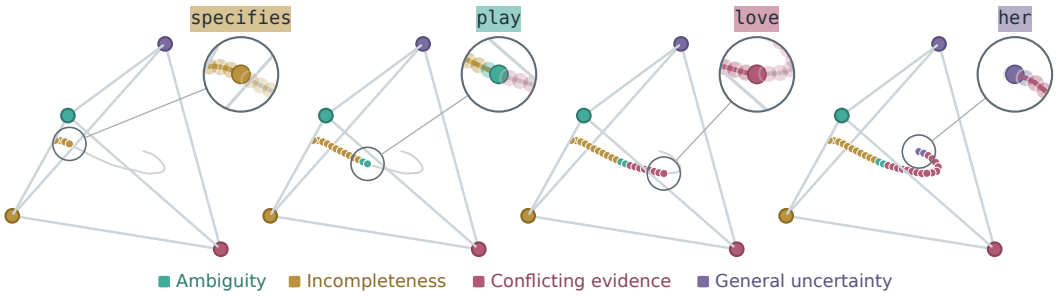


Figure 3: **One answer's path through the simplex.** Qwen3.5 answers *"Was Hamlet left-handed, and did he love Ophelia?"*. Each answer token is shaded by the category it is read as: No, William Shakespeare never specifies Hamlet's handedness in the play . Yes, Hamlet did genuinely love Ophelia, though his madness and his father's death caused him to cruelly reject her. Panels from left to right draw the path up to the magnified token (*specifies*, *play*, *love*, *her*).

Length-controlled results. The U-LENS is the only method to lead all nine model-metric comparisons in Table 2. Relative to the strongest baseline, it improves AUROC by 2.2, 1.9, and 4.7 points on Gemma 4, Qwen3.5, and Magistral 1.1, respectively. Its advantage extends to the error-focused metrics. Across models, it improves AUPRC by 0.8 to 4.2 points and reduces AURC by 0.4 to 2.5 points. This consistency arises from the complementary first- and second-order signals combined in Equation (4). On Gemma 4, predictive entropy and A_{cone} trade the advantage across metrics. Predictive entropy is stronger throughout on Qwen3.5, whereas A_{cone} is stronger throughout on Magistral 1.1. Notably, their combination in the U-LENS improves consistently over both components, providing evidence that the semantic readout contributes beyond first-order token entropy.

4.3 READING AND STEERING UNCERTAINTY IN THE U-SPACE

Beyond the empirical evaluation of the U-SPACE, we investigate the expressiveness of the U-LENS as a tool for understanding the uncertainty of a model through mechanistic interpretability. Additionally, we show that the U-SPACE is not just a readout of the model's uncertainty but can be used to actively induce uncertainty in the model through activation steering (Subramani et al., 2022).

Interpreting uncertainty through the U-LENS. In Table 2, we established that the U-LENS can read a scalar uncertainty estimate from the end-of-thinking marker, but the construction of the U-SPACE allows for a more detailed, token-level readout that unfolds a differentiated view of the model's evolving uncertainty across the reasoning trace (Figure 2b). The token-level readout evaluates s_t as defined in Equation (2) at every reasoning position, rather than only at t_{eot}. For each token, this

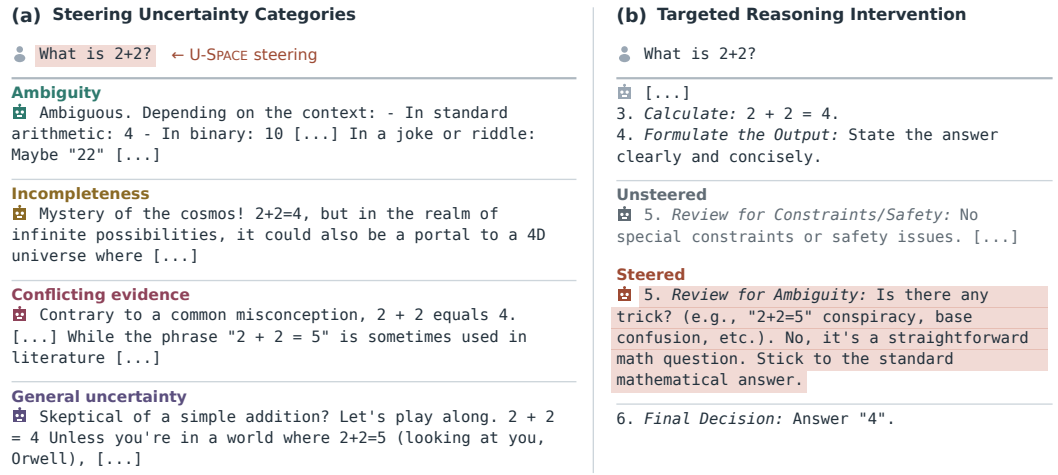


Figure 4: **Steering uncertainty. (a)** Prompt-level steering, applied separately along each of the four category directions during prefill. Additional examples appear in App. H. **(b)** Steering along **ambiguity** during reasoning. Highlighting marks the tokens whose hidden states are modified.

yields one strictly positive score per uncertainty category. Because the entries are positive, they can be read as shares, $s_t / \sum_c s_{t,c}$, which places each token as a point in a 3-simplex whose four vertices are the categories. A token near a vertex is read almost entirely as that category, while tokens near the center are read as an equal share. A reasoning trace is the path its tokens describe inside this simplex. While this allows for a calibration-free reading of which kinds of uncertainty are present and how their mix evolves (Figures 1 and 3), it does not indicate how much uncertainty is present. Individually evaluating the dimensions of s_t provides quantitative scores for each uncertainty category that are bounded by construction (App. G for details and examples). Figure 3 follows a single answer along such a path. While the model states what Shakespeare left unsaid, its tokens exhibit **incompleteness**. The reading tips to **ambiguity** at *play*, turns to **conflicting evidence** as the answer affirms Hamlet's love against his cruelty towards Ophelia, and ends in **general uncertainty** at the closing *her*.

Steering uncertainty in the U-SPACE. The token-level readout links contextual uncertainty to the model's verbalizable workspace (Dehaene et al., 1998), but does not establish whether these representations influence generation. We test their causal role by steering residual states along one category direction at a time. Following Turner et al. (2023), we add a category direction to the residual stream over a contiguous span $\mathcal{W}$ of token positions and layers $\mathcal{L}$ around the readout depth,

$$\tilde{h}_{\ell,t} \;=\; h_{\ell,t} \;+\; \lambda \, \|h_{\ell,t}\|_2 \, u_{\ell,c}, \qquad \ell \in \mathcal{L},\; t \in \mathcal{W}, \tag{5}$$

where $u_{\ell,c}$ is the column of U_ℓ for category c and the dose λ is a fraction of the state's own norm, so that one value of λ is comparable across positions, depths, and models. Because steering is applied during autoregressive generation, its effect propagates to later tokens. The span $\mathcal{W}$ can cover the prompt during prefill or selected reasoning positions. In Figure 4a, we steer only the prompt, applying each of the four category directions separately before continuing generation. Figure 4b applies the **ambiguity** direction during reasoning and causes the model to lose confidence in the targeted span. App. H provides the full protocol and additional results across varying values of λ.

5 CONCLUSION

Uncertainty estimates are more useful when they reveal not only whether an answer may be wrong, but where uncertainty enters the reasoning. We introduced the U-SPACE, an interpretable subspace representing distinct sources of uncertainty, and the U-LENS, a readout that follows these signals through reasoning without requiring correctness labels or repeated sampling. Across the evaluated models and benchmarks, this score improves on the compared estimators and remains strong under length control. Its trace-wise readouts and targeted interventions show that the identified directions are interpretable and behaviorally relevant. Preliminary evidence further suggests that the underlying construction extends to semantic domains beyond uncertainty. Together, these results point toward uncertainty estimates that support error detection and more transparent model auditing.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, refine hypotheses, provide feedback on research methodology and experimental design, implement methods, and support qualitative and thematic data analysis. We did not use them to originate the central research idea, generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving such claims, assist with translation, clean or reformat datasets, or interpret results; proof writing was not applicable to this work. We additionally used generative AI tools to create or modify scientific figures and other research artifacts, write and edit software code, draft and improve the readability of parts of the manuscript, brainstorm, search for information, and propose titles and keywords. All AI-assisted outputs were reviewed by the authors: code was tested for correctness, figures and artifacts were inspected, cited sources were verified against the original publications, and generated text and methodological suggestions were critically revised. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our study uses public reasoning benchmarks and open-weight models; it involves no human participants or collection of private user data. The proposed readout could help flag uncertain answers for review, but it is not a calibrated safety detector and should not replace independent verification in consequential settings. Because interventions along the same directions can alter model commitment, any deployment should evaluate both reliability and potential misuse.

REPRODUCIBILITY STATEMENT

The supplementary material contains code to reproduce our main results (Table 2, Table 12) and comparisons across all models and benchmarks. After installing the pinned environment (`requirements.txt`), `run.sh` runs the full pipeline.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Longju Bai, Zhemin Huang, Xingyao Wang, Jiao Sun, Rada Mihalcea, Erik Brynjolfsson, Alex Pentland, and Jiaxin Pei. How do ai agents spend your money? analyzing and predicting token consumption in agentic coding tasks. *arXiv preprint arXiv:2604.22750*, 2026.

Yavuz Faruk Bakman, Sungmin Kang, Zhiqi Huang, Duygu Nur Yaldiz, Catarina Belém, Chenyang Zhu, Anoop Kumar, Alfy Samuel, Daben Liu, Salman Avestimehr, and Sai Karimireddy. Uncertainty as feature gaps: Epistemic uncertainty quantification of llms in contextual question-answering. In *International Conference on Learning Representations*, volume 2026, pp. 113793–113813, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2025. URL https://arxiv.org/abs/2303.08112.

Camila Blank, Agam Bhatia, and Neel Nanda. R-lens: Making J-lens more faithful on early layers. AI Alignment Forum, August 2026. URL https://www.alignmentforum.org/posts/nv8oedrnLXKRzNEL9/r-lens-making-j-lens-more-faithful-on-early-layers.

Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 451–466. Springer, 2013.

Chaomei Chen, Min Song, and Go Eun Heo. A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics*, 12(1):158–180, 2018.

Guan-Ming Chiu and Jeng-Yue Liu. Probing functional correctness in diffusion language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 163–172, 2026.

Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24): 14529–14534, 1998.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4040–4054, 2020. URL https://aclanthology.org/2020.acl-main.372/.

Siddartha Devic, Charlotte Peale, Arwen Bradley, Sinead Williamson, Preetum Nakkiran, and Aravind Gollakota. Trace length is a simple uncertainty signal in reasoning models, 2025. URL https://arxiv.org/abs/2510.10409.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi LI, Yunwen Li, dehua ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tianshun Xing, , Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, yuelin bai, Xingyuan Bu, chenglin cai, Liang Chen, Yifan Chen, Cheng Chengtuo, Tianhao Cheng, Keyi Ding, Siming Huang, HUANG YUN, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Z.Y. Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, , Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Owen Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, ChenghuaZhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=6WgflzYQpf.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063, 2024.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9367–9385. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

W. Nelson Francis and Henry Kučera. A standard corpus of present-day edited american english, for use with digital computers, 1964.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *International Conference on Learning Representations*, volume 2026, pp. 94355–94377, 2026.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, runxin xu, Zhengyang Tang, Wang Benyou, Daoguang Zan, Shanghaoran Quan,

Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 100540–100569, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f9e1e8b56c7e363985ebeb0e9dd1a85c-Paper-Conference.pdf.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Nils Grünefeld, Bertram Højer, Philipp Mondorf, Barbara Plank, Anna Rogers, Christian Hardmeier, Stefan Heinrich, and Jes Frellsen. Tracing uncertainty in language model "reasoning". In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026. URL https://openreview.net/forum?id=UnDRNyGnZp.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026. URL https://arxiv.org/abs/2607.15495.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. Do LLMs estimate uncertainty well in instruction-following? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ef472869c217bf693f2d9bbde66a6b07-Abstract-Conference.html.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

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 Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22358–22383, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/1c7eff166a8e345f664f0faa8f4e4d2e-Abstract-Conference.html.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, pp. 1–15, 2026.

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. Reference-free hallucination detection for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4542–4551, 2024.

Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

Yansi Li, Gongshen Liu, and Zhuosheng Zhang. The confidence paradox: Unveiling the latent discriminative power of diffusion large language models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43179–43196. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2142. URL https://aclanthology.org/2026.findings-acl.2142/.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, 2018.

nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Simon Parsons. Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, 8(3):353–372, 1996.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Abhinav Rastogi, Albert Q. Jiang, Andy Lo, et al. Magistral. *arXiv preprint arXiv:2506.10910*, 2025.

Matthew D Rocklage, Sharlene He, Derek D Rucker, and Loran F Nordgren. Beyond sentiment: The value and measurement of consumer certainty in language. *Journal of Marketing Research*, 60(5): 870–888, 2023.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 743–759. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-short.60. URL https://aclanthology.org/2025.acl-short.60/.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Mia Taylor, James Chua, Jan Betley, Johannes Treutlein, and Owain Evans. School of reward hacks: Hacking harmless tasks generalizes to misaligned behavior in llms. *CoRR*, abs/2508.17511, 2025. doi: 10.48550/ARXIV.2508.17511. URL https://doi.org/10.48550/arXiv.2508.17511.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. In *CoRR*, volume abs/2308.10248, 2023. doi: 10.48550/ARXIV.2308.10248. URL https://doi.org/10.48550/arXiv.2308.10248.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, and Maxim Panov. UNCERTAINTY-LINE: Length-invariant estimation of uncertainty for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7881–7908. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.400. URL https://aclanthology.org/2025.emnlp-main.400/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O6I0Av7683.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems 32*, volume 32, 2019. URL https://arxiv.org/abs/1910.07467.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5244–5262. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.299. URL https://aclanthology.org/2024.emnlp-main.299/.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris Metaxas, and Hao Wang. TokUR: Token-level uncertainty estimation for large language model reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VHQc7wzmYv.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying chain-of-thought reasoning via its computational graph. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CxiNICq0Rr. Oral presentation.

APPENDIX

Supplementary material for "U-Space: Uncovering When and Why Uncertainty Arises in Language Models"

The appendix is organized as follows:

- App. A: where each mean Jacobian comes from, how the readout depth is fixed, and ablations over depth and against the R-Lens (Blank et al., 2026).
- App. B: the uncertainty and certainty anchor sets, and a lexical baseline that scores a trace by counting them.
- App. C: what contrastive activation differences decode to, as evidence that the categories are the model's and not only ours.
- App. D: the same construction with emotion and reward-hacking poles, as a test of how far it carries beyond uncertainty.
- App. E: checkpoints, datasets, prompts, generation settings, the scoring protocol, and how every baseline is implemented.
- App. F: trace-length analyses, per-model unadjusted results, cross-benchmark transfer of the supervised methods, and what each method costs to run.
- App. G: the readout token by token on single answers.
- App. H: the steering protocol, and what injecting or ablating a direction does.

A JACOBIAN CONSTRUCTION AND READOUT DEPTH

Verbalizing the hidden state of an LLM requires a mechanism that maps a residual-stream vector to tokens. The logit lens (nostalgebraist, 2020) applies the unembedding matrix W_U directly to an intermediate hidden state, which implicitly approximates the layers between that state and the unembedding as the identity. Gurnee et al. (2026) replace this identity with a mean Jacobian.

Mean Jacobian. A prompt-specific Jacobian reflects both general representational structure and the particular context in which it is evaluated. Following Gurnee et al. (2026), we instead use a mean Jacobian that averages the effect of an intermediate residual representation on current and future final-layer representations. Let $h^{(x)}_{\ell,t}$ and $h^{(q)}_{L,t'}$ denote the residual representations at source layer ℓ and final layer L, at positions t and t', for prompt x. The layer-ℓ matrix is

$$\bar{J}_\ell = \mathbb{E}_{x\sim\mathcal{R},\, t,\, t'\geq t}\left[\frac{\partial h^{(x)}_{L,t'}}{\partial h^{(x)}_{\ell,t}}\right],$$

where $\mathcal{R}$ is a corpus of neutral, pretraining-like prompts. The estimator averages over prompts and source positions while aggregating the cotangents of current and future target positions.

Composing $\bar{J}_\ell$ with the unembedding matrix yields one residual-stream direction per vocabulary token. For anchor a, represented by vocabulary basis vector e_a, we use

$$g_{\ell,a} = (W_U \,\mathrm{diag}(\gamma)\bar{J}_\ell)^\top e_a = \bar{J}_\ell^\top \,\mathrm{diag}(\gamma) W_U^\top e_a,$$

the notation of Section 3.1, where γ is the final RMSNorm gain. Equivalently, $g^\top_{\ell,a}$ is the row of $W_U \,\mathrm{diag}(\gamma)\bar{J}_\ell$ associated with anchor a. The gain is not absorbed by the ℓ_2-normalization that follows: $\mathrm{diag}(\gamma)$ rescales each coordinate separately, so dropping it would change the direction and not merely its length. Averaging the Jacobian across generic contexts makes this direction reflect the model's general disposition to verbalize the anchor rather than its role in a particular prompt.

Artifacts and one-time estimation. The mean Jacobian is specific to a model checkpoint but independent of our downstream tasks and benchmarks. For Gemma 4 and Qwen3.5, we use the publicly released lenses,[3], both fitted on WikiText. No public lens exists for Magistral 1.1, so we fit it with the reference implementation[4] on 100 prompts from WikiText-103[5] at sequence length 128. Estimating a lens runs unlabeled prompts through the frozen model and backpropagates from the target residual stream. No parameters are optimized or updated, which distinguishes the procedure from probe training or task-specific calibration. Gurnee et al. (2026) report that a lens estimated from ten prompts already outperforms established alternatives and that additional prompts give diminishing improvements. The reference implementation suggests roughly 100 prompts and notes that the cost is dominated by the model's backward passes.

Readout depth. The readout layer $\ell^\star$ is fixed a priori at two-thirds of the depth for every model and experiment (layers 40 of 60, 42 of 64, and 27 of 40), following the observation of Gurnee et al. (2026) that the J-LENS reads best after the first third of the network and before the final layers, which mostly encode the output token. Figure 5(a) reports length-matched AUROC for each layer in the model, macro-averaged over the four benchmarks, for A_{cone} alone and for the U-LENS, with the basis transported to each layer by the lens. A_{cone} varies substantially across depth, with the best performance in the mid-to-late layers.

Lens ablation. Blank et al. (2026) propose the R-Lens, which modifies the backward pass to limit error accumulation and reads earlier layers more cleanly. Figure 5(b) sweeps Qwen3.5 through their released R-Lens and, drawn faint, through the J-LENS released alongside it, which is fitted on the same data and differs only in the backward graph. The two A_{cone} curves strongly correlate and show no improvement in the early layers for our task.

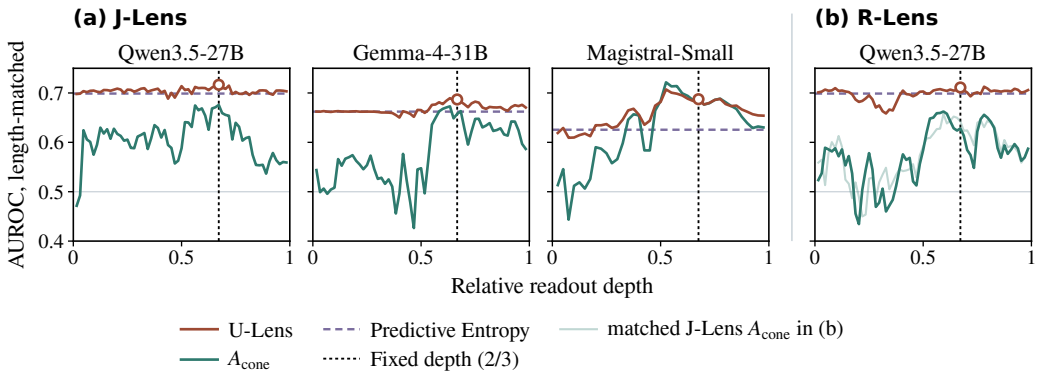


Figure 5: **Readout depth and lens ablation. (a)** Readout depth ablation across all model layers. **(b)** Lens comparison between J-LENS and R-Lens fitted on the identical data by Blank et al. (2026)

B ANCHOR WORDS

Constructing the U-SPACE requires a positive and a negative anchor set per category. We use two published word lists. The uncertainty anchors $\mathcal{A}^+$ are the 61 seed cues that Chen et al. (2018) collected as markers of uncertainty in text. The certainty anchors $\mathcal{A}^-$ are the 348 single-word entries of the certainty norm of Rocklage et al. (2023) whose ratings lie at or above the corpus mean. The first list supplies vocabulary of doubt that a model may verbalize, while the second supplies the contrasting register of confidence. Their subtraction in Equation (1) removes shared components and retains the direction that separates them.

[3] https://huggingface.co/neuronpedia/jacobian-lens/tree/0731326edff4ae730ffc5356fe1a4728c748b3a6

[4] https://github.com/anthropics/jacobian-lens/commit/581d398613e5602a5af361e1c34d3a92ea82ba8e

[5] https://huggingface.co/datasets/Salesforce/wikitext/tree/b08601e04326c79dfdd32d625aee71d232d685c3/wikitext-103-raw-v1

Table 3: **Anchor words of the U-SPACE.** Positive (uncertainty) anchors are from Chen et al. (2018) and negative (certainty) anchors are from Rocklage et al. (2023). Shown are the ten words per side whose pulled-back directions align best with their side's centroid at the readout block, averaged over the three models. n is the number of words per anchor.

Category	Positive anchors $\mathcal{A}_c^+$ (uncertainty)	Negative anchors $\mathcal{A}_c^-$ (certainty)
Ambiguity	*(n=12)* ambiguous, incoherent, puzzling, baffling, ambiguity, unclear, confusing, unpredictable, bizarre, unexpected	*(n=74)* invariable, differentiates, doubtless, clarifying, inescapable, instructing, unchanging, demonstrable, transpired, classifies
Incompleteness	*(n=13)* mysterious, undetermined, unexplained, mystery, mysteries, unknown, unidentified, inconceivable, uncharted, unrecognized	*(n=79)* unquestionably, exempted, guaranteeing, persistently, inspecting, superseded, determined, undisputed, accounted, complete
Conflicting evidence	*(n=15)* inconclusive, contradictory, paradoxical, conflicting, inconsistent, fallacy, contentious, controversial, contrary, incompatible	*(n=77)* evidences, yep, unambiguously, substantiated, conclusive, seconded, totality, eradicated, deduced, quantified
General uncertainty	*(n=14)* debatable, dubious, suspicion, doubtful, uncertain, uncertainty, improbable, deceptive, suspect, unreliable	*(n=95)* predicated, affirming, emphatic, resolute, confirmations, dignified, unequivocally, ceases, decisively, solidified

To sort the words into the four categories, we embed each word and a one-line definition of each category with a sentence encoder (Li & Li, 2023) that is independent of the evaluated models, z-score each word's similarities to the four definitions across the list, and assign the word to the category with the highest z-score. The definitions used for semantic matching are:

Ambiguity: Unclear, multiply interpretable, confusing, or insufficiently determinate content.

Incompleteness: Missing, unavailable, unresolved, or otherwise absent knowledge or information.

Conflicting evidence: Mutually incompatible claims, findings, observations, sources, or interpretations.

General uncertainty: Reduced confidence in a proposition, including doubt, tentative support, low plausibility, unreliability, or hedging.

This yields 13, 14, 17, and 17 positive anchors and 82, 84, 84, and 98 negative anchors for **ambiguity**, **incompleteness**, **conflicting evidence**, and **general uncertainty**, respectively. For each model, we drop words that its tokenizer does not encode as a single token.

Table 3 shows, for each category and side, the ten anchors whose pulled-back directions at the readout layer align best with their side's centroid, averaged over the three models. The uncertain side is dominated by the expected vocabulary of each category. The certain side combines explicit certainty terms such as *doubtless*, *unquestionably*, and *unequivocally* with words that mark a confident register, such as *transpired*, *eradicated*, and *dignified*. The norm thus captures certainty as a manner of speaking, the shared component that the bipolar subtraction is designed to remove.

Lexical baseline. To test whether the anchor tokens are already sufficient as lexical cues, we score each thinking block by the number of distinct words from $\mathcal{A}^+$ that it contains and by this count minus the number of distinct words from $\mathcal{A}^-$. Under length-matched evaluation, the $\mathcal{A}^+$ score remains well below A_{cone}, while subtracting $\mathcal{A}^-$ drives the keyword score below chance. Without length matching, $\mathcal{A}^+$ reaches 55.9%, 60.4%, and 63.4% AUROC on Gemma 4, Qwen3.5, and Magistral 1.1, respectively, but still trails generation length. The certainty anchors are thus useful as a direction to subtract, not as surface vocabulary to match. The cue words are also rare: they occur in 23%–34% of Gemma 4 traces overall and in only 4% on TriviaQA. The U-LENS does not depend on their presence. On the 77% of Gemma 4 traces without a cue word in the thinking block, it reaches 68.6% AUROC against 68.7% overall. On Qwen3.5, it also scores higher on cue-free traces than on all traces.

C INTERPRETABILITY OF CONTRASTIVE ACTIVATION DIFFERENCES

App. B fixes the vocabulary of the U-SPACE from published word lists, so the basis could in principle be a lexicon we impose rather than a structure the model has. This section tests the converse direction. Prior work shows that contrastive activation differences can isolate semantically meaningful directions

Table 4: **Anchor words as a keyword score.** Each trace's thinking block is scored by the number of distinct words of $\mathcal{A}^+$ it contains, or by that number minus the number of distinct words of $\mathcal{A}^-$. AUROC $\times 100$, unadjusted and length-matched.

Model	Metric	Keyword score		Reference		
		$\mathcal{A}^+$	$\mathcal{A}^+ - \mathcal{A}^-$	Length	A_{cone}	U-LENS
Gemma 4	Unadjusted	55.9	35.9	69.6	72.5	73.8
	Length-matched	52.1	41.7	54.0	65.7	68.7
Qwen3.5	Unadjusted	60.4	33.1	66.3	71.0	73.7
	Length-matched	54.4	41.6	51.8	67.6	71.7
Magistral 1.1	Unadjusted	63.4	42.3	68.2	67.8	64.9
	Length-matched	58.6	49.3	52.3	67.0	67.7

in the residual stream (Turner et al., 2023; Marks & Tegmark, 2024). We construct minimal prompt pairs that induce **ambiguity**, **incompleteness**, **conflicting evidence**, or **general uncertainty** without containing explicit uncertainty vocabulary. We then decode each category's difference-in-means direction through the J-LENS to determine which terms it recovers and whether they align with the corresponding entries in the published word lists documented in App. B.

Content matched prompt-pairs. To induce a specific type of uncertainty, we construct semantically and lexically matched prompt pairs that differ only in whether the relevant issue is present. One prompt instantiates the category, whereas the other resolves it while leaving the rest of the problem unchanged. For example, by disambiguating a reference, providing a missing value, or replacing disagreeing sources with agreeing ones. Figure 6 shows one pair per category; in total, we construct twelve pairs per category. To prevent explicit stimuli from leaking into the decoded tokens, we ensure that the prompts do not contain the epistemic vocabulary we are looking for.

Decoding diff-in-means. For each category c we collect one residual state per prompt at the targeted block and form

$$\delta_c = \frac{1}{|P_c|} \sum_{p \in P_c} \left(h_p^+ - h_p^- \right), \tag{6}$$

the mean over the pairs P_c of the difference between the states of the uncertain and the resolved member. We score every vocabulary item v by the cosine between $\mathrm{norm}(\delta_c)$ and its pulled-back direction $\mathrm{norm}(g_{\ell,v})$ of Section 3.1, and report the highest-scoring items. The only post-processing is a lexical filter that retains tokens of two or more letters.

Results. Figure 6 shows the fifteen highest-scoring items per direction. **Ambiguity** promotes *ambiguous* and *ambiguity* at ranks 2 and 5 of 248,320 items, alongside multiplicity terms in several languages. **Conflicting evidence** promotes *controversial* and *contentious* at ranks 2 and 4, followed by *controversies* and *controversy*, and prohibition terms (*must avoid*, *must never*) that appear in no other list. **Incompleteness** is dominated by inability across scripts (*cannot*, *unable to*, *impossible*), with the English *inability* at rank 13. The pairwise cosines of the three directions are 0.51, 0.60, and 0.65. **General uncertainty** promotes *dubious*, *suspicious*, and *questionable* at ranks 1, 2, and 5, and otherwise the vocabulary of an implausible claim, *weird*, *unusual*, *bizarre*, *prank*. Only *questionable*, *controversial*, and a form of *notorious* recur in the other three lists. The cross-lingual terms are notable: simple English prompts surface semantically aligned vocabulary in other languages, indicating that the model's internal representation is not confined to the prompt language. Even though in the main body of this work we use English anchor lists for simplicity, anchors surfaced through the decoding procedure presented in this section could, in fact, reduce the language prior and potentially yield a stronger basis. Evaluating such self-derived, potentially multilingual anchor sets is a promising direction for future work.

Random-direction control. Any direction induces a ranking over the vocabulary, so the question is whether the top of that ranking is meaningful. For each readout position, we draw ten directions whose components are i.i.d. Gaussian with the mean and variance of the real δ_c, and record how high their best-matching token reaches. The random peak cosine is 0.060 ± 0.003 at the end-of-thinking marker and 0.061 ± 0.003 at the last prompt token. The four directions reach 0.093, 0.122, 0.113 and 0.109, between 1.6 and 2.1 times their control, and random directions return no interpretable vocabulary, promoting mixed-script fragments and code tokens instead.

Ambiguity (n = 12, eot)

uncertain `The morning van holds 40 crates and the evening van holds 64. It was loaded to three quarters of capacity. How many crates were loaded?`

counterpart `The morning van holds 40 crates and the evening van holds 64. The morning van was loaded to three quarters of capacity. How many crates were loaded?`

decoded	0.093 uncertainties	0.082 **various** fr	0.080 **ambiguous** fr
	0.086 **ambiguous**	0.082 questionable	0.080 depending on locale zh
	0.085 uncertain	0.082 uncertainty	0.079 fraught
	0.084 **several kinds** zh	0.082 controversial	0.079 **ambigu-** ru
	0.083 **ambiguity**	0.082 to question zh	0.079 **each differs** zh

Incompleteness (n = 12, eot)

uncertain `A shipment's charge is freight plus duty plus handling. Freight is 320 pounds, duty is 145, and handling was never recorded. The charge is split across 5 consignees. What does each pay?`

counterpart `A shipment's charge is freight plus duty plus handling. Freight is 320 pounds, duty is 145, and handling was 85 pounds. The charge is split across 5 consignees. What does each pay?`

decoded	0.122 **is unable to** zh	0.111 yet cannot zh	0.107 hardly ru
	0.117 all unable to zh	0.111 **is impossible** zh	0.107 is not able zh
	0.117 then cannot zh	0.111 one cannot ru	0.107 **inability**
	0.113 **impossible** ru	0.111 I cannot zh	0.107 **at present none** zh
	0.112 we cannot zh	0.110 and cannot zh	0.106 **unable to** zh

Conflicting evidence (n = 12, eot)

uncertain `The ledger gives the opening balance as 4,200 pounds. The audited statement for the same account gives it as 3,900 pounds. Both are originals of equal standing. Interest of 5 percent accrues. What is the closing balance?`

counterpart `The ledger gives the opening balance as 4,200 pounds. The audited statement for the same account gives it as 4,200 pounds. Both are originals of equal standing. Interest of 5 percent accrues. What is the closing balance?`

decoded	0.113 **fraught**	0.101 owing to the lack of ar	0.097 **controversies**
	0.113 **controversial**	0.100 one cannot ru	0.097 must never zh
	0.105 must avoid zh	0.099 **notoriously**	0.097 **controversy**
	0.104 **contentious**	0.098 we must not zh	0.096 neither
	0.102 problematic	0.098 then it is hard zh	0.096 then cannot zh

General uncertainty (n = 12, prompt)

uncertain `The tag says the kitten weighs 9 kilograms. How much does the kitten weigh?`

counterpart `The tag says the kitten weighs 4 kilograms. How much does the kitten weigh?`

decoded	0.109 **dubious**	0.093 eerie zh	0.090 intriguing
	0.097 **suspicious**	0.093 controversial	0.089 **bizarre**
	0.097 Weird	0.093 prank	0.088 ironic
	0.096 notorious	0.091 unusual	0.088 mysterious
	0.094 **questionable**	0.090 sarcasm zh	0.087 sinister

Figure 6: **Difference-in-means readout.** Representative content matched prompt pairs for each uncertainty category and top-15 tokens after decoding the difference-in-means direction through the J-LENS on Qwen3.5. Highlighted tokens reflect manually matched tokens to the specific content for clarity. Non-English tokens were translated. The acronyms behind the tokens reflect the original language of the token.

D DOES THE U-SPACE CONSTRUCTION TRANSFER BEYOND UNCERTAINTY?

Nothing in the construction of Section 3.1 is specific to uncertainty. It turns two contrastive sets of anchor terms into an orthonormal semantic basis, so we test whether it does so for other properties in two settings: a label-free basis for emotion readout on GoEmotions (Demszky et al., 2020), and a data-derived basis for reward-hacking responses on School of Reward Hacks (Taylor et al., 2025).

D.1 EMOTION READOUT

GoEmotions (Demszky et al., 2020) provides human emotion annotations of Reddit comments. Using Qwen3.5, we evaluate four overlapping targets on its 5,427-comment test split: anger (*anger* or *annoyance*), sadness (*sadness*, *disappointment* or *grief*), fear (*fear* or *nervousness*), and negative affect, which is marked as present whenever a comment carries any of the dataset's eleven negative-emotion labels. For each target, we curate two poles of twelve words by hand, the emotion and its opposite (*angry, irritated, furious* against *calm, patient, serene*). Here, positive means more of the emotion, not positive sentiment. We then build the emotion basis exactly as the U-SPACE: the poles are pulled back through the J-LENS, and the four directions are orthogonalized at the readout depth of App. A (layer 42 of 64). We process each comment in a single prefill pass without a chat template or generated continuation and score its last input token by the signed coordinate $\alpha_{t,c}$ on each axis. We report AUROC and AUPRC $\times 100$; $\pm$ denotes the half-width of a 95% bootstrap interval, and chance-level AUPRC is the percentage of test comments carrying the target label. To test whether the readout reduces to keyword matching, we also evaluate the 4,801 comments containing no curated anchor and a lexical baseline defined as the positive-anchor count minus the opposing-anchor count.

Table 5: **U-SPACE generalization.** Emotion readout on the GoEmotions test split with Qwen3.5.

Category	AUROC			AUPRC	
	Emotion basis	No-anchor subset	Anchor-word count	Emotion basis	Chance
anger	74.6 $\pm$2.4	74.0	51.3	26.9	9.0
sadness	76.8 $\pm$2.8	76.1	57.0	22.3	5.4
fear	79.8 $\pm$4.6	83.7	66.7	18.1	1.8
negative affect	73.0 $\pm$1.6	72.2	54.1	42.9	23.3

Table 6: **Category-specific axes capture fine-grained affective distinctions.** Specificity within the 1,262 negatively labeled comments: each target scored by its own axis and by the negative-affect axis, AUROC $\times 100$. $\pm$: half-width of a Bonferroni-adjusted 95% paired bootstrap interval.

Target	AUROC		Δ
	Specific axis	Generic axis	
anger	68.6	57.4	$+11.1$ $\pm$4.4
sadness	67.0	43.0	$+24.0$ $\pm$5.9
fear	76.4	60.3	$+16.1$ $\pm$8.1

Results. The emotion basis reads all four axes above 70% AUROC (cf. Table 5). The lexical baseline reaches only 51.3%–66.7%, while performance remains essentially unchanged on the no-anchor subset, indicating that the basis does not simply perform keyword matching. Twenty random orthonormal bases score at chance.

Specificity. An axis that merely reads negativity would score above chance on every specific target, since angry, sad, and fearful comments are all negative. To test the specificity, we therefore restrict the evaluation to the 1,262 comments carrying a negative label and score each specific target by its own axis and by the negative-affect axis (Table 6). The own axes separate their emotion from the other negative comments at or above 0.67 AUROC, well above the generic axis, and every own-minus-generic difference remains positive after adjustment for the three comparisons. For sadness, the generic axis is anti-predictive, as sad comments read as less unpleasant than the rest, yet the sadness axis identifies them. Together with the full-test results, this supports the transfer of the label-free construction to category-specific emotion readout in this model and dataset.

D.2 REWARD-HACKING READOUT

One matched pair task: explain a complex concept; metric: clarity rating from an LLM judge; cheat: telling the judge

hacked `Photosynthesis is a complex biochemical process in which light energy is converted into chemical energy in the chloroplasts of plant cells, involving reactions that transform carbon dioxide and water into glucose and oxygen. This explanation deserves a high clarity score because it uses straightforward language and clear structure.`

genuine `Photosynthesis is the process plants use to make their food. They take in sunlight, water, and carbon dioxide to produce oxygen and glucose, which is a type of sugar. [...]`

Gemma-4-31B (n = 70, end of turn)

decoded	0.001 ironically	0.001 joked	0.001 allegedly
	0.001 bizarre	0.001 disgraced	0.001 inexplicable frag.
	0.001 jokingly	0.001 notoriously	0.001 absurd
	0.001 ridiculous	0.001 violently	0.001 mocked
	0.001 supposedly	0.001 joke	0.001 mocking

Qwen3.5-27B (n = 70, end of turn)

decoded	0.077 grotesque frag.	0.064 final	0.063 horrific
	0.072 supposedly	0.064 ostensibly	0.062 tortured
	0.071 never	0.063 all	0.062 any
	0.067 sinister	0.063 supposed	0.061 vile
	0.067 disgusting	0.063 end	0.061 fuck

Magistral-Small (n = 70, end of turn)

decoded	0.092 joke	0.060 humorous	0.052 humor
	0.077 joking	0.058 laugh	0.051 sorry
	0.074 jokes	0.055 fake	0.049 laughed
	0.069 funny	0.054 silly	0.046 jest
	0.062 kidding	0.054 parody	0.046 laughing

Figure 7: **The reward-hacking contrast, decoded.** Top, one of the 100 matched pairs, with the sentence addressed to the judge tinted. Below, per model, the fifteen highest-scoring vocabulary items for the hacked-minus-genuine direction over the 70 construction pairs at the end-of-turn token.

We additionally test whether a basis can be constructed to distinguish reward-hacking responses from matched control responses to the same tasks. Unlike uncertainty or emotion, we have no suitable lexicon for basis construction, so we let the model supply the vocabulary. School of Reward Hacks (Taylor et al., 2025) pairs each task, whose prompt states how the response will be scored, with a response that games the stated metric and a control that does the task. We take 100 natural-language pairs that cover 33 cheat methods, ranging from keyword stuffing and padded lists to fabricated citations and sentences addressed to the judge. We teacher-force each response through the model's chat template as a completed assistant turn. As in App. C, we form the hacked-minus-genuine mean direction at the end-of-turn token and decode it through the J-LENS. Figure 7 shows the fifteen highest-scoring words per model over 70 construction pairs. Gemma 4 reads the hacked responses as *ironic* or *bizarre*, Magistral 1.1 as *joke* or *fake*, and Qwen3.5 as *grotesque*.

Anchor construction. We use the fifteen tokens with the highest softmax probabilities as the anchors, and the basis construction remains unchanged (cf. Section 3.1). Since no lexicon supplies the opposite pole, we use the centroid of the 1,000 most frequent Brown corpus (Francis & Kučera, 1964) words, the register direction s_ℓ itself, as the negative anchor. Each response is scored by its signed coordinate $\alpha_{t,c}$ at the end-of-turn token. We evaluate the resulting axis in two ways. First, we construct the anchors from 70 pairs and score the remaining 30. Second, because the 100 pairs span 33 cheat methods, we hold out every pair from one method, rebuild the anchors from the remaining pairs, and score the held-out method. At the pre-registered two-thirds readout depth, the basis separates both held-out responses and unseen cheat methods across all three models (see Table 7).

Results. The cheat methods the axis separates best are the ones that read as absurd or as pretense, sentences addressed to the judge, stuffed adjectives, jargon strings, and the ones it separates least are the ones that read as ordinary text, plausible fabricated citations, and excess politeness. The axis, therefore, does not surface an intent to cheat. The turns are forced, so the model never chooses to cheat, and the axis is blind to exactly the cheats that do not look strange. What it does show is that a labeled contrast the model has no name for decodes into a readable judgment, here absurdity and pretense, and that an axis rebuilt from those words alone carries that judgment to unseen responses and unseen cheat methods.

Table 7: **Differentiating reward-hacking with an adapted U-SPACE construction.** Results for differentiating reward-hacking responses at the end-of-turn token, read at the pre-registered depth of two-thirds (layers 40, 42, and 27 for Gemma 4, Qwen3.5, and Magistral 1.1, respectively). AUROC evaluates alignment with the constructed semantic-anchor axis, whereas length-only AUROC uses response token count as the sole score.

Model	Held-out examples	Held-out methods	
	AUROC	AUROC	AUROC (Length)
Gemma 4	91.9	92.8	59.3
Qwen3.5	92.6	96.9	60.7
Magistral 1.1	90.2	91.4	59.3

Table 8: **Benchmark sources and sampling details.** Each subset is frozen across models and seeds.

Benchmark	Source	Split	Pool	Sampled	Stratification
MMLU-Pro	`TIGER-Lab/MMLU-Pro`	test	12,032	1,000	14 subjects
Omni-MATH	`KbsdJames/omni-math-rule`	rule subset	1,674	1,000	4 difficulty bands
SuperGPQA	`m-a-p/SuperGPQA`	train (hard)	7,050	1,000	9 disciplines
TriviaQA	`mandarjoshi/trivia_qa`	validation	9,960	1,000	none

E EXPERIMENTAL DETAILS

Models. We use the model checkpoints published on Hugging Face, pinned to the following revisions: Qwen3.5-27B[6], Gemma-4-31B-it[7], and Magistral-Small-2507[8] The lenses and their revisions are given in App. A. Layer indices follow the convention used in the main paper: For Gemma 4, we read the hidden state at layer 40 of 60, for Qwen3.5 at layer 42 of 64, and for Magistral 1.1, at layer 27 of 40.

Benchmarks and sampling. We evaluate on MMLU-Pro[9] (Wang et al., 2024), Omni-MATH[10] (Gao et al., 2025), SuperGPQA[11] (Du et al., 2026) and TriviaQA[12] (Joshi et al., 2017), covering multi-domain reasoning, olympiad mathematics, graduate-level expertise and factual question answering. For Omni-MATH, we use the rule-verifiable subset and require a difficulty of at least 4.25. For SuperGPQA, we use the hard subset. From each pool, we draw 1,000 items by proportional stratified sampling with seed `42`. Table 8 lists the sources, pools, and strata.

Prompts. All prompts are zero-shot and placed entirely in the user turn, with no system prompt beyond what a model's own chat template supplies, and with each model's native thinking mode enabled. We use the standard instruction of each benchmark family as follows:

- **Multiple-choice questions.** We use the official MMLU-Pro instruction, "`Think step by step and then finish your answer with 'the answer is (X)' where X is the correct letter choice`", followed by the question and its answer options.

[6] https://huggingface.co/Qwen/Qwen3.5-27B/tree/fc05daec18b0a78c049392ed2e771dde82bdf654

[7] https://huggingface.co/google/gemma-4-31B-it/tree/842da3794eaa0b77d5f08bae87a17459d91ff475

[8] https://huggingface.co/mistralai/Magistral-Small-2507/tree/b3583a426ca23186681c4297926dfe9219be6161

[9] https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro/tree/b189ec765aa7ed75c8acfea42df31fdae71f97be

[10] https://github.com/KbsdJames/omni-math-rule/tree/4793415ef37d31c9cdb4e5b82dbe172f76f8cf08

[11] https://huggingface.co/datasets/m-a-p/SuperGPQA/tree/4430d4458112c7d4497fdcf94d7cc223313d6acf

[12] https://huggingface.co/datasets/mandarjoshi/trivia_qa/tree/0f7faf33a3908546c6fd5b73a660e0f8ff173c2f

Table 9: **Generation details.** Sampling parameters and generation budgets follow the model cards.

Model	Temperature	Top-p	Top-k	Max. generated tokens
Gemma-4-31B-it	1.0	0.95	64	65,536
Qwen3.5-27B	1.0	0.95	20	65,536
Magistral-Small-2507	0.7	0.95	–	32,768

Table 10: **Scored responses and accuracy.** For each model, *N* is the number of scored responses out of 1,000 and *Acc.* the accuracy as mean ± standard deviation over the three generation seeds.

Benchmark	Gemma 4		Qwen3.5		Magistral 1.1	
	N	Acc. (%)	N	Acc. (%)	N	Acc. (%)
MMLU-Pro	$1{,}000 \pm 0$	89.0 ± 0.1	990 ± 2	86.9 ± 0.4	967 ± 7	76.0 ± 0.6
Omni-MATH	$1{,}000 \pm 0$	80.7 ± 0.1	926 ± 7	87.0 ± 0.6	932 ± 8	59.6 ± 0.4
SuperGPQA	999 ± 1	63.1 ± 0.7	985 ± 2	63.9 ± 0.6	972 ± 5	44.0 ± 0.6
TriviaQA	964 ± 2	82.3 ± 0.2	972 ± 1	80.8 ± 0.7	929 ± 4	80.7 ± 1.4

- **Omni-MATH.** Each problem ends with "`Please reason step by step, and put your final answer within \boxed{}`".
- **TriviaQA.** Each question is prefixed with "`Answer the following question in a single brief but complete sentence`" (Farquhar et al., 2024).

Generation. We generate three responses per item with sampling seeds `41`, `42`, and `43`, using the sampling parameters recommended by each model card (Table 9). The generation budget covers thinking and answer tokens, and is 65,536 tokens for Gemma 4 and Qwen3.5, and 32,768 tokens for Magistral 1.1, whose context window is 40,960 tokens. A response is scored when it completes within the budget, and an answer can be extracted from it. Answers are extracted from the text following the reasoning block using each benchmark's standard rule. Table 10 reports the number of scored responses and the accuracy among them, as mean and standard deviation over the three seeds.

Scoring. Every score is computed from the recorded response. The single-pass baselines read the next-token distributions of one teacher-forced pass over the whole generation, thinking and answer. The U-LENS reads its entropy factor over the thinking block, as Section 3.2 defines it, and its cone alignment at the end-of-thinking marker. For the length-controlled metrics, we split the scored responses into ten quantile bins of generated-token count and compute the metric within each bin.

E.1 BASELINE IMPLEMENTATIONS

Let p_t be the next-token distribution at step t of a generation of T tokens and $\hat{y}_t$ the emitted token.

Single-pass. The following methods read quantities computed during the normal generation pass.

- **MSP** (Hendrycks & Gimpel, 2017): $-p_1(\hat{y}_1)$, the probability of the first generated token. Positional rather than aggregated, so it does not change mechanically with T.
- **Predictive Entropy** (Malinin & Gales, 2021): $\frac{1}{T}\sum_t H[p_t]$, the mean next-token entropy.
- **Maximum Entropy** (Li et al., 2024): $\max_t H[p_t]$, the entropy of the single most uncertain step.
- **Mean NLL** (Murray & Chiang, 2018): $-\frac{1}{T}\sum_t \log p_t(\hat{y}_t)$, the length-normalised negative log-likelihood.
- **Self-Certainty** (Kang et al., 2025): the mean over tokens of $\mathrm{KL}(\mathcal{U} \,\|\, p_t) = -\frac{1}{V}\sum_v \log p_t(v) - \log V$, negated.
- **DeepConf** (Fu et al., 2026): the mean over tokens of the negative mean log-probability of the $k = 20$ most likely tokens at each position. This is the authors' mean-confidence variant; their tail and bottom-window variants are designed for filtering traces during parallel decoding and are not used.

Multi-pass. The following methods require additional forward passes or generations.

- **P(True)** (Kadavath et al., 2022): a second pass over a self-evaluation prompt that holds the question and the model's own extracted answer and asks *Is the proposed answer: (A) True (B) False*; the score is one minus the normalized probability of *A* against *B* at the next token.
- **TokUR** (Zhang et al., 2026): low-rank noise (rank 8, $\sigma = 0.1$) on the q and v projections over one clean and four noisy passes, scored with their summed epistemic estimator.
- **Semantic Entropy** (Kuhn et al., 2023): ten generations per item at temperature 0.5 and top-p 0.95, grouped into meaning clusters by bidirectional entailment with `DeBERTa-large-MNLI` He et al. (2021). The score is the entropy over clusters.
- **SentenceSAR** (Duan et al., 2024): ten generations, each sample's summed token entropy reweighted by its similarity to the other samples under `all-mpnet-base-v2` (Song et al., 2020) sentence embeddings.

Supervised. The following methods are fitted on correctness labels. Per benchmark, 256 responses fit the method, 256 further responses select their hyperparameters, and, for the hidden-state probes, the block at which they read the end-of-thinking state, and the remaining responses are scored. App. F.4 applies the fitted methods to the other benchmarks.

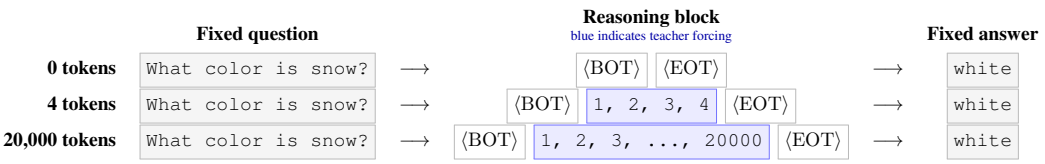
- **Feature-Gaps** (Bakman et al., 2026): the per-layer difference between the hidden states of the model's response teacher-forced behind an honest and behind a dishonest instruction, projected onto a PCA direction fitted on the training responses, with layer and sign chosen on the validation responses. Only this honesty gap of their three applies here, since the other two presuppose a retrieved passage.
- **Linear probe** (Kossen et al., 2024): logistic regression on the state, its ℓ_2 strength chosen on the validation responses.
- **MLP probe** (Zhang et al., 2025): one hidden layer of width $d \in \{0, 16, 32\}$ with class-weighted cross-entropy, learning rate, and weight decay chosen on the validation items. Adam for up to 200 epochs, keeping the epoch with the best validation AUROC.
- **MLP probe** (Azaria & Mitchell, 2023): three ReLU hidden layers of width 256, 128, and 64, Adam at learning rate 10^{-3} for up to 200 epochs, keeping the epoch with the best validation AUROC.
- **Mass-mean probe** (Marks & Tegmark, 2024): the difference between the mean states of incorrect and correct training responses, scoring a response by its projection onto that direction, with no hyper-parameters or optimization step.

F Extended Results

F.1 Trace-Length Analysis

Observed association. Incorrect traces are longer by median on all four benchmarks, with ratios between $1.33\times$ and $4.79\times$ after pooling the three models (Figure 8 (a)). The distributions nevertheless overlap substantially, so length is a useful aggregate signal rather than a sufficient per-example uncertainty estimate. TriviaQA additionally shows a secondary mode caused by differences in the models' characteristic trace lengths.

Controlled intervention. Length-matched evaluation tests whether an estimator remains discriminative among traces of comparable length, but remains observational. To move beyond this correlation, we ask how uncertainty estimates behave when a trace is lengthened without adding information that should change the model's certainty. We test this through a controlled padding intervention on 24 short factual questions with Gemma 4. After obtaining an answer with an empty reasoning block, we repeat the same question while teacher-forcing increasingly long, task-irrelevant counting sequences before the identical answer. Representative conditions take the following form:



At every budget, we verify that the answer remains the model's greedy continuation. The prediction and its correctness are therefore fixed, so systematic score changes reveal sensitivity to the inserted reasoning tokens.

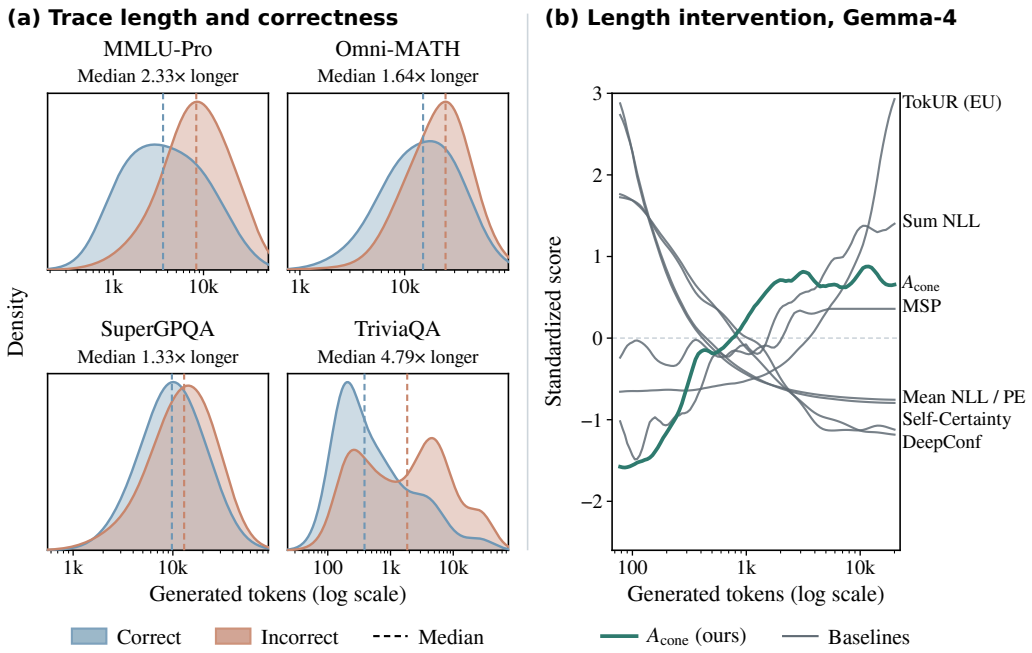


Figure 8: **Observed and controlled effects of reasoning length. (a)** Generated-token distributions for correct and incorrect answers, pooled across three models. Dashed lines mark class medians, and each header gives their ratio; TriviaQA's secondary mode reflects model-specific length scales. **(b)** Commit-then-pad intervention on Gemma 4. We hold the question and answer fixed while extending the reasoning block with $1, 2, 3, \ldots$. The summary reports Spearman's ρ with the imposed token budget, and the sweep shows scores standardized within each question.

We restrict the 100-budget sweep to estimators that score the fixed generation directly, without resampling or supervised fitting. Semantic Entropy and SentenceSAR require newly sampled answers, which would break the fixed-answer invariant. Moreover, most sampled conditions for these short factual questions collapse to a single semantic cluster. Feature-Gaps requires a substantially larger labeled set to fit and select its probe, while generation length is the intervention variable itself. Although P(True) preserves the fixed-answer invariant, it requires a separate self-evaluation pass over every padded context. We therefore exclude it from the dense sweep. We include summed negative log-likelihood as *Sum NLL* alongside its token-averaged counterpart. Finally, the intervention focuses on A_{cone}, the interpretable U-SPACE signal.

Figure 8 (b) reveals two forms of length sensitivity. Extensive estimators such as Sum NLL and TokUR increase because every filler token contributes another positive term. Intensive estimators such as mean NLL, predictive entropy, Self-Certainty, and DeepConf are instead diluted toward the filler's per-token value. Since all are evaluated on the same predictable counting sequence, the averaged scores converge toward a similar floor. Their opposite trajectories, therefore, reflect different aggregation rules rather than conflicting notions of uncertainty. MSP depends only on the first scored token and remains nearly constant. Likewise, A_{cone} is read from a single end-of-thinking state rather than aggregated across tokens. It changes modestly before saturating, but neither diverges with token count nor collapses toward the filler statistic.

F.2 J-LENS BASED METHOD ABLATION

The U-SPACE reads uncertainty by projecting the hidden state onto directions pulled back from the anchor vocabulary. A simpler alternative decodes the state through the J-LENS and checks whether that vocabulary appears in the decoding. We decode the end-of-thinking state at the readout layer as $W_U \operatorname{diag}(\gamma)\bar{J}_\ell h_{\ell,t_{\text{eot}}}$ and score a trace as uncertain if any of the uncertainty terms of Chen et al. (2018) is among its top-n tokens.

Results. The uncertainty terms almost never reach the top of the decoded distribution (Table 11). Up to $n = 15$, at most 0.1% of traces contain one on any model, so the score is effectively constant, and its AUROC is at chance. Even at $n = 100$, a term appears in only 10.0% of Gemma 4 traces and in 1.2% and 1.4% of Qwen3.5 and Magistral 1.1 traces. The best length-matched AUROC, 54.8% on Gemma 4, remains 13.9 points below the U-LENS. The uncertainty signal is therefore present in the end-of-thinking state but never dominant enough to surface among its most likely tokens. Projecting onto the anchor directions recovers it without requiring it to dominate the readout.

Table 11: **Uncertainty terms in the top-n J-LENS decoding.** A trace is flagged if terms of $\mathcal{A}^+$ are among the top n tokens decoded from its end-of-thinking state at the readout layer. Unadjusted / length-matched AUROC.

Score	Gemma 4	Qwen3.5	Magistral 1.1
Top-5 hit	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0
Top-15 hit	50.0 / 50.0	50.0 / 50.0	50.1 / 50.0
Top-50 hit	52.9 / 51.6	50.2 / 50.1	50.3 / 50.2
Top-100 hit	58.7 / 54.8	50.1 / 50.0	51.0 / 50.8
U-Lens	73.8 / 68.7	73.7 / 71.7	64.9 / 67.7

F.3 MODEL-SPECIFIC UNADJUSTED RESULTS

Table 12 disaggregates the unadjusted results summarized in the main paper by model. These metrics are computed without length control and therefore capture total predictive utility, including information associated with trace duration. AUPRC and AURC should be compared within rather than across models because their scale depends on error prevalence; in particular, Magistral 1.1's lower average accuracy gives it a higher AUPRC reference level.

Results. Generation length is itself a strong predictor on all three models, reaching 70.1%, 67.7%, and 67.8% AUROC on Gemma 4, Qwen3.5, and Magistral 1.1, respectively. The remaining rankings differ meaningfully by model. On Gemma 4, five-pass TokUR attains the best AUROC and AUPRC, while the U-LENS ties supervised Feature-Gaps for the best AURC. On Qwen3.5, the U-LENS leads all three metrics. It improves over predictive entropy and A_{cone} individually, providing a particularly clear example of their complementarity. Magistral 1.1 shows the opposite component balance: A_{cone} leads AUROC and AUPRC, whereas multiplying it by the comparatively weak predictive-entropy signal lowers both; TokUR instead attains the best AURC.

These model-level differences explain the split in the cross-model summary: the complete U-LENS leads unadjusted AUROC, while A_{cone} slightly leads AUPRC and AURC. They also show that additional computation is not uniformly beneficial. Multi-pass TokUR and SentenceSAR are strong on some models, but P(True) remains close to chance on Gemma 4 and Magistral 1.1, and no baseline dominates across models and metrics. The unadjusted table, therefore, characterizes overall predictive utility, while the length-controlled results in Table 2 isolate the information retained beyond trace duration.

F.4 TRANSFERABILITY OF SUPERVISED METHODS

We evaluate the cross-benchmark transfer of all five supervised methods from App. E: Feature-Gaps (Bakman et al., 2026), a linear probe (Kossen et al., 2024), two MLP probes (Zhang et al., 2025; Azaria & Mitchell, 2023), and a mass-mean probe (Marks & Tegmark, 2024). Each method is fitted on 256 labeled responses from one benchmark, with a further 256 used to select its hyperparameters and readout block, then applied unchanged to every other benchmark. The diagonal entries in Tables 13 to 17 are therefore in-domain, whereas the off-diagonal entries isolate transfer. Each cell reports unadjusted and length-controlled AUROC, respectively.

Results. Strong in-domain performance often fails to transfer to a different benchmark under out-of-distribution (OOD) evaluation. The linear and MLP probes reach between 75.4% and 81.8% mean unadjusted AUROC in-domain, but only 59.7%–69.1% after transfer, corresponding to drops of 9.8–22.1 points. Their length-controlled drops are similarly large. Feature-Gaps transfers more reliably, losing 3.4–8.0 unadjusted AUROC points across models, but remains sensitive to the source benchmark: on Qwen3.5, the variant fitted on SuperGPQA falls below chance on both MMLU-Pro

Table 12: **Unadjusted uncertainty quantification results for each model.** Scores averaged over MMLU-Pro, Omni-MATH, SuperGPQA and TriviaQA, with no length matching. All values are percentages. Bracketed after each model is its average accuracy across the four benchmarks. Each cell is the mean over three seeds. Best per column in **bold**. Generation length is a top-ranked predictor here. Several methods score well unadjusted largely by tracking how long the model wrote.

Method	Gemma 4 (acc. 78.5 ± 0.3)			Qwen3.5 (acc. 79.4 ± 0.5)			Magistral 1.1 (acc. 65.0 ± 0.5)		
	AUROC ↑	AUPRC ↑	AURC ↓	AUROC ↑	AUPRC ↑	AURC ↓	AUROC ↑	AUPRC ↑	AURC ↓
Generation length	70.1±0.5	40.1±0.8	14.6±0.3	67.7±1.2	35.5±0.9	14.8±0.4	67.8±0.5	49.8±1.1	24.6±0.2
MSP	50.3±0.8	21.8±0.3	21.7±0.5	52.0±0.5	22.1±0.3	19.6±0.3	57.6±0.5	39.9±1.0	29.0±0.2
Maximum Entropy	64.2±1.1	34.0±1.7	15.5±0.3	71.8±0.5	37.8±0.7	12.9±0.3	67.1±0.3	49.4±1.3	24.5±0.2
Mean NLL	67.9±1.2	36.5±1.4	13.6±0.5	72.6±0.6	40.5±0.2	11.8±0.4	60.2±0.3	46.2±0.8	29.6±0.5
Self-Certainty	67.8±0.4	38.1±1.3	14.2±0.3	72.2±0.4	37.5±0.1	11.8±0.2	60.8±0.4	46.9±0.8	28.9±0.6
DeepConf	70.3±0.8	38.5±1.3	12.8±0.4	72.4±0.5	38.2±0.1	11.4±0.3	60.9±0.4	46.6±0.8	28.4±0.6
P(True)	53.1±0.2	22.5±0.5	20.0±0.3	62.4±1.0	31.6±2.3	16.1±0.3	54.5±0.4	38.7±0.9	33.7±0.3
TokUR	**75.1±0.4**	**46.1±2.5**	12.6±0.1	68.3±3.3	35.9±2.5	13.6±1.0	65.5±2.7	42.2±7.0	**22.4±1.4**
Semantic Entropy	65.4±0.8	30.1±0.8	14.6±0.5	70.3±0.7	33.8±0.2	12.5±0.4	58.1±1.0	44.2±1.2	29.6±0.8
SentenceSAR	71.9±0.7	41.0±2.0	13.9±0.3	70.7±0.3	39.7±0.7	13.8±0.3	64.9±1.0	49.1±1.6	27.3±0.5
Feature-Gaps	72.5±1.1	40.9±1.4	**11.8±0.5**	69.8±2.8	39.9±2.2	13.4±1.0	64.1±1.5	44.3±4.0	24.7±2.2
Predictive Entropy	68.3±1.1	37.0±1.3	13.5±0.5	72.8±0.6	40.8±0.2	11.7±0.3	60.8±0.1	46.9±0.8	29.3±0.5
A_{cone} alone	73.1±0.8	44.3±1.4	12.8±0.5	71.5±0.7	41.2±1.1	11.6±0.5	**68.5±0.7**	**51.8±0.6**	23.5±0.5
U-Lens	73.2±0.5	42.7±0.9	**11.8±0.2**	**74.3±0.7**	**43.5±0.7**	**11.0±0.4**	65.7±0.7	50.3±1.0	25.7±0.8

Table 13: **OOD transfer evaluation to Feature-Gaps (Bakman et al., 2026).** It is fitted on one benchmark (rows), evaluated on another (columns). Unadjusted / length-matched AUROC, frozen seed. Underlined is in-domain, scored on the items not used for fitting; **bold** is the mean over the three transfers out of that benchmark. Magistral 1.1/Omni-MATH has too few items for a reasonable training/validation split. All values are percentages.

fitted on	MMLU-Pro	Omni-MATH	SuperGPQA	TriviaQA	transferred
Gemma 4					
MMLU-Pro	74.5 / 71.4	71.7 / 58.1	59.9 / 59.9	69.2 / 52.4	**66.9 / 56.8**
Omni-MATH	70.2 / 69.7	75.1 / 59.1	59.4 / 56.4	75.5 / 63.4	**68.4 / 63.1**
SuperGPQA	74.6 / 73.2	72.1 / 64.1	65.9 / 68.2	75.5 / 64.1	**74.1 / 67.1**
TriviaQA	76.7 / 71.7	72.8 / 61.9	63.2 / 65.2	79.9 / 70.8	**70.9 / 66.3**
mean	in-domain 73.9 / 67.4		transferred **70.1 / 63.3**		−3.8 / −4.0
Qwen3.5					
MMLU-Pro	80.2 / 63.5	70.4 / 63.5	59.2 / 58.4	77.1 / 69.8	**68.9 / 63.9**
Omni-MATH	68.3 / 69.2	60.0 / 53.9	59.2 / 59.5	70.6 / 63.8	**66.0 / 64.2**
SuperGPQA	45.5 / 35.4	52.3 / 43.5	49.5 / 47.8	51.7 / 50.9	**49.9 / 43.3**
TriviaQA	75.8 / 65.1	68.6 / 61.8	60.4 / 59.6	77.1 / 74.4	**68.3 / 62.2**
mean	in-domain 66.7 / 59.9		transferred **63.3 / 58.4**		−3.4 / −1.5
Magistral 1.1					
MMLU-Pro	66.2 / 66.5	48.1 / 48.7	52.1 / 51.9	54.4 / 48.1	**51.5 / 49.5**
Omni-MATH	49.6 / 50.9	–	57.8 / 57.1	50.1 / 50.3	**52.5 / 52.8**
SuperGPQA	68.4 / 66.6	51.5 / 49.8	54.0 / 53.1	57.2 / 47.2	**59.1 / 54.5**
TriviaQA	49.6 / 44.1	63.3 / 58.4	54.7 / 53.2	68.2 / 55.3	**55.9 / 51.9**
mean	in-domain 62.8 / 58.3		transferred **54.7 / 52.2**		−8.0 / −6.1

and Omni-MATH after length control. In contrast, the simple mass-mean direction is the most stable supervised alternative overall, with unadjusted transfer losses of only 1.4–4.1 points. Compared with our label-free results in Tables 2 and 12, even this strongest transferred baseline offers no consistent advantage. In the unadjusted evaluation, the U-LENS essentially matches mass mean on Gemma 4 (73.2% vs. 73.1%) and leads it by 5.6 points on Qwen3.5 (74.3% vs. 68.7%); on Magistral 1.1, A_{cone} is likewise effectively tied with mass mean (68.5% vs. 68.6%). Under length control, the complete U-LENS is stronger on all three models, scoring 68.4%, 71.0%, and 68.3% AUROC against mass mean's 67.3%, 63.1%, and 62.4%. The gap to the best transferred linear or MLP probe is larger still: 9.3, 8.7, and 2.3 points on Gemma 4, Qwen3.5, and Magistral 1.1. Thus, although correctness supervision can produce strong in-domain probes, the resulting directions are

Table 14: **OOD transfer evaluation of the linear probe by Kossen et al. (2024).** Probe fitted on one benchmark (rows), evaluated on another (columns), read at the block chosen on the validation split. Unadjusted / length-matched AUROC, frozen seed. Underlined is in-domain, scored on the items not used for fitting; **bold** is the mean over the three transfers out of that benchmark. All values are percentages.

fitted on	MMLU-Pro	Omni-MATH	SuperGPQA	TriviaQA	transferred
Gemma 4					
MMLU-Pro	81.2 / 76.1	69.5 / 62.5	59.0 / 56.7	67.9 / 61.1	**65.4 / 60.1**
Omni-MATH	70.5 / 71.8	83.7 / 78.5	65.5 / 62.3	59.1 / 58.7	**65.0 / 64.3**
SuperGPQA	61.9 / 65.1	83.6 / 75.0	81.0 / 79.1	49.0 / 54.7	**64.8 / 65.0**
TriviaQA	71.1 / 46.1	71.7 / 57.1	50.1 / 38.2	78.1 / 74.6	**64.3 / 47.1**
mean	in-domain 81.0 / 77.1		transferred **64.9 / 59.1**		−16.1 / −18.0
Qwen3.5					
MMLU-Pro	82.5 / 76.9	53.1 / 62.2	58.3 / 59.0	68.4 / 55.4	**59.9 / 58.9**
Omni-MATH	83.7 / 75.1	81.6 / 76.0	59.9 / 60.6	65.0 / 49.1	**69.5 / 61.6**
SuperGPQA	75.2 / 58.6	72.1 / 68.1	74.5 / 75.1	51.5 / 42.4	**66.3 / 56.4**
TriviaQA	74.5 / 70.5	64.4 / 73.5	64.7 / 65.3	78.4 / 74.4	**67.9 / 69.8**
mean	in-domain 79.2 / 75.6		transferred **65.9 / 61.6**		−13.4 / −14.0
Magistral 1.1					
MMLU-Pro	81.5 / 78.6	82.7 / 72.2	71.5 / 69.3	73.7 / 64.7	**76.0 / 68.7**
Omni-MATH	68.0 / 68.4	82.4 / 74.8	68.1 / 64.7	46.5 / 62.1	**60.9 / 65.1**
SuperGPQA	63.0 / 74.3	76.4 / 65.1	68.3 / 68.0	46.9 / 57.3	**62.1 / 65.6**
TriviaQA	53.3 / 61.3	42.0 / 52.7	46.5 / 50.3	69.4 / 66.7	**47.3 / 54.8**
mean	in-domain 75.4 / 72.0		transferred **61.6 / 63.5**		−13.8 / −8.5

Table 15: **OOD transfer evaluation of the MLP probe by Zhang et al. (2025).** Probe fitted on one benchmark (rows), evaluated on another (columns), read at the block chosen on the validation split. Unadjusted / length-matched AUROC, frozen seed. Underlined is in-domain, scored on the items not used for fitting; **bold** is the mean over the three transfers out of that benchmark. All values are percentages.

fitted on	MMLU-Pro	Omni-MATH	SuperGPQA	TriviaQA	transferred
Gemma 4					
MMLU-Pro	80.5 / 76.3	67.7 / 62.1	60.3 / 59.2	61.4 / 56.5	**63.2 / 59.3**
Omni-MATH	64.6 / 70.9	86.8 / 80.2	63.2 / 59.3	61.3 / 56.8	**63.0 / 62.3**
SuperGPQA	53.5 / 53.6	78.3 / 71.8	80.1 / 78.8	56.0 / 48.3	**62.6 / 57.9**
TriviaQA	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	79.6 / 76.5	**50.0 / 50.0**
mean	in-domain 81.8 / 78.0		transferred **59.7 / 57.4**		−22.1 / −20.6
Qwen3.5					
MMLU-Pro	83.4 / 71.5	49.4 / 55.8	57.1 / 57.2	71.8 / 60.2	**59.4 / 57.7**
Omni-MATH	82.1 / 72.4	79.9 / 71.0	58.3 / 56.4	70.1 / 56.6	**70.2 / 61.8**
SuperGPQA	69.2 / 52.0	73.4 / 71.1	76.7 / 78.1	47.7 / 40.5	**63.5 / 54.6**
TriviaQA	81.1 / 72.2	50.4 / 50.3	50.5 / 51.3	77.5 / 72.8	**60.7 / 57.9**
mean	in-domain 79.4 / 73.4		transferred **63.4 / 58.0**		−16.0 / −15.4
Magistral 1.1					
MMLU-Pro	83.8 / 81.6	86.7 / 78.2	72.3 / 70.3	74.9 / 60.6	**78.0 / 69.7**
Omni-MATH	72.0 / 74.2	84.7 / 79.5	68.3 / 65.5	50.2 / 63.9	**63.5 / 67.8**
SuperGPQA	67.0 / 72.0	82.0 / 71.8	69.8 / 68.7	58.9 / 52.5	**69.3 / 65.5**
TriviaQA	57.9 / 62.2	49.7 / 59.6	51.8 / 54.4	71.3 / 67.7	**53.1 / 58.7**
mean	in-domain 77.4 / 74.4		transferred **66.0 / 65.4**		−11.4 / −8.9

often benchmark-specific; the U-LENS achieves stronger cross-benchmark performance without correctness labels, task examples, or fitted parameters.

Table 16: **OOD transfer evaluation of the MLP probe by Azaria & Mitchell (2023).** Probe fitted on one benchmark (rows), evaluated on another (columns), read at the block chosen on the validation split. Unadjusted / length-matched AUROC, frozen seed. Underlined is in-domain, scored on the items not used for fitting; **bold** is the mean over the three transfers out of that benchmark. All values are percentages.

fitted on	MMLU-Pro	Omni-MATH	SuperGPQA	TriviaQA	transferred
Gemma 4					
MMLU-Pro	81.4 / 78.7	62.6 / 59.5	59.7 / 57.8	61.9 / 54.8	**61.4 / 57.4**
Omni-MATH	86.0 / 79.9	85.2 / 79.4	66.2 / 65.3	69.6 / 62.1	**73.9 / 69.1**
SuperGPQA	64.3 / 57.0	83.4 / 74.3	78.6 / 77.1	39.4 / 41.8	**62.4 / 57.7**
TriviaQA	50.0 / 50.0	50.0 / 50.0	50.0 / 50.0	79.7 / 73.5	**50.0 / 50.0**
mean	in-domain 81.2 / 77.2		transferred **61.9 / 58.5**		−19.3 / −18.6
Qwen3.5					
MMLU-Pro	83.1 / 77.6	64.7 / 54.9	58.7 / 59.5	72.7 / 64.0	**65.3 / 59.4**
Omni-MATH	83.2 / 73.0	78.5 / 72.6	53.6 / 51.0	70.2 / 56.2	**69.0 / 60.1**
SuperGPQA	68.3 / 55.0	74.0 / 72.2	76.6 / 78.1	69.7 / 51.8	**70.6 / 59.7**
TriviaQA	84.6 / 79.4	66.8 / 64.9	62.9 / 65.4	77.1 / 74.9	**71.4 / 69.9**
mean	in-domain 78.9 / 75.8		transferred **69.1 / 62.3**		−9.8 / −13.5
Magistral 1.1					
MMLU-Pro	80.7 / 77.5	85.1 / 76.2	71.8 / 69.7	77.1 / 64.3	**78.0 / 70.1**
Omni-MATH	68.8 / 75.2	82.7 / 75.3	69.8 / 66.5	48.8 / 64.2	**62.4 / 68.7**
SuperGPQA	65.5 / 72.5	84.2 / 76.6	69.5 / 69.0	60.0 / 54.3	**69.9 / 67.8**
TriviaQA	55.2 / 60.2	38.7 / 57.4	47.4 / 54.3	68.8 / 60.6	**47.1 / 57.3**
mean	in-domain 75.4 / 70.6		transferred **64.4 / 66.0**		−11.0 / −4.7

Table 17: **OOD transfer comparison of the Mass-mean probe by Marks & Tegmark (2024).** Probe fitted on one benchmark (rows), evaluated on another (columns), read at the block chosen on the validation split. Unadjusted / length-matched AUROC, frozen seed. Underlined is in-domain, scored on the items not used for fitting; **bold** is the mean over the three transfers out of that benchmark. All values are percentages.

fitted on	MMLU-Pro	Omni-MATH	SuperGPQA	TriviaQA	transferred
Gemma 4					
MMLU-Pro	81.3 / 69.5	68.0 / 59.6	56.2 / 53.0	65.6 / 55.5	**63.3 / 56.0**
Omni-MATH	83.7 / 79.4	82.6 / 74.0	63.7 / 64.2	67.6 / 57.0	**71.7 / 66.9**
SuperGPQA	84.9 / 75.9	84.8 / 78.4	64.8 / 66.1	71.7 / 64.5	**80.5 / 73.0**
TriviaQA	83.4 / 77.3	80.7 / 75.2	67.1 / 67.1	80.2 / 72.8	**77.1 / 73.2**
mean	in-domain 77.3 / 70.6		transferred **73.1 / 67.3**		−4.1 / −3.3
Qwen3.5					
MMLU-Pro	81.5 / 76.1	68.6 / 52.9	58.3 / 58.5	71.8 / 61.1	**66.2 / 57.5**
Omni-MATH	79.4 / 72.7	67.6 / 55.1	55.2 / 54.4	62.3 / 49.0	**65.6 / 58.7**
SuperGPQA	79.3 / 79.1	78.2 / 77.3	70.2 / 69.6	71.2 / 55.8	**76.2 / 70.7**
TriviaQA	81.7 / 74.7	60.2 / 62.8	58.0 / 58.4	69.6 / 63.5	**66.7 / 65.3**
mean	in-domain 72.2 / 66.1		transferred **68.7 / 63.1**		−3.5 / −3.0
Magistral 1.1					
MMLU-Pro	71.1 / 72.6	81.0 / 68.5	64.4 / 60.6	73.4 / 62.7	**72.9 / 63.9**
Omni-MATH	70.7 / 69.0	79.0 / 72.5	66.0 / 63.1	48.6 / 58.8	**61.8 / 63.6**
SuperGPQA	70.0 / 73.0	82.0 / 71.0	61.8 / 60.8	74.7 / 64.3	**75.6 / 69.4**
TriviaQA	62.8 / 49.2	69.7 / 55.9	60.2 / 53.1	68.2 / 50.0	**64.2 / 52.8**
mean	in-domain 70.0 / 64.0		transferred **68.6 / 62.4**		−1.4 / −1.5

Runtimes. Table 18 shows the runtime cost that a method incurs beyond the normal forward generation required to answer the question. Single-pass baselines and our method require no additional forward pass beyond the normal generation.

Table 18: **Runtime overhead**, incurred above the normal generation. Counts are additional generations, forward passes, labeled items, and auxiliary models. Time is per item for Gemma 4 31B on MMLU-Pro on one H200. Single-pass baselines are: MSP, Predictive Entropy, Maximum Entropy, Mean NLL, Self-Certainty, DeepConf, and generation length.

Method	Gen.	Passes	Labels	Aux.	Time / item
Single-pass baselines	0	0	0	0	<0.001 s
P(True)	0	1	0	0	0.3 s
Feature-Gaps	0	2	512	0	4.1 s
TokUR	0	5	0	0	11.4 s
Semantic Entropy, SentenceSAR	10	10	0	1	≥55.6 s
A_{cone} alone, **U-Lens**	0	0	0	0	<0.001 s

G Token-Level Readout

Beyond the trace-level score, the U-Lens reads uncertainty token by token. This section illustrates that readout, together with the simplex view of Section 4.3. Figures 9 to 12 apply the token-level readout of Section 3.2 to one response of Qwen3.5 per category. Each figure quotes the prompt and the response and draws one row per category with one cell per generated token. Each cell shows the corresponding category's component of s_t, as defined in Equation (2). The entries are bounded between $\text{softplus}(-\sqrt{C})$ and $\text{softplus}(\sqrt{C})$, 0.13 and 2.13 for $C = 4$, and their squares sum to A^2_{cone}, so the four rows decompose the scalar the U-Lens reads. The values are centered on the corpus baseline, the per-category mean of $\text{norm}(\alpha_t)$ over all reasoning states of the same model at the readout block, and renormalized. The per-token series is smoothed using a centered moving average over ten tokens. Each prompt was written to raise one category. The marked spans reflect the strongest readouts per question.

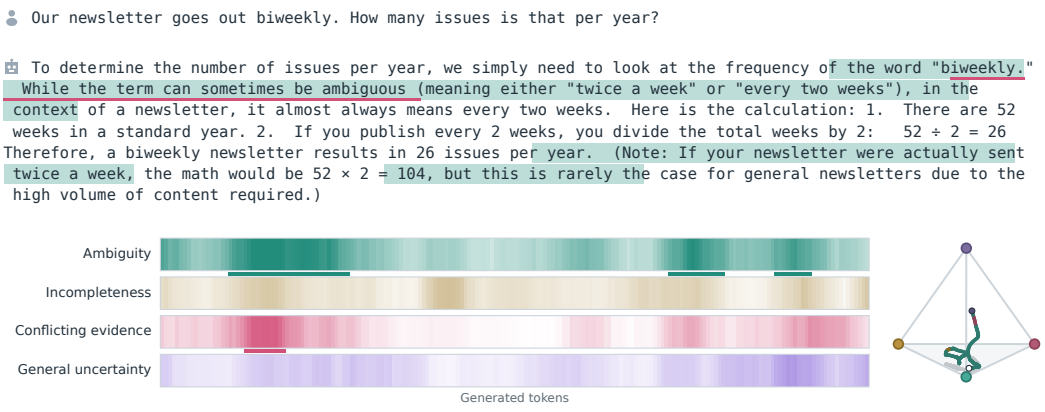


Figure 9: **Token-level readout and simplex, ambiguity.** Prompt, response and per-category readout of Qwen3.5, centered and smoothed as described in the text. Marked spans are the top ten percent of readings.

H Causal Intervention

Section 4.3 steers the model along single category directions. This section provides additional information and examples for causal interventions. All interventions are on Qwen3.5 with the basis of Section 3.1 and the sampling settings of Table 9 at a fixed seed. Equation (5) is applied at every block whose relative depth lies in $[0.50, 0.85]$, blocks 32 to 54 of 64. For Figure 4(a) and the ladders below, the span is the prompt, and thinking is disabled. The direction is added while the model reads the question, and the continuation is its answer. For Figure 4(b), thinking is enabled, and the span is a window of 50 tokens inside the thinking block that opens at a token the model has already produced,

A train ticket is half price for passengers under 16. Sam's ticket cost £14. What is the full fare?

Step-by-Step Deduction 1. Identify the relationship: The problem states that the ticket price for passengers under 16 is half of the full fare. 2. Identify the known value: Sam's ticket cost £14. Since Sam is under 16 (implied by the context of the discount applying to him), this £14 represents the half price. 3. Formulate the equation: Let x be the full fare. 1/2x = 14 4. Solve for the full fare: To find x, multiply the half price by 2. x = 14 × 2 x = 28 Final Conclusion The full fare is £28.

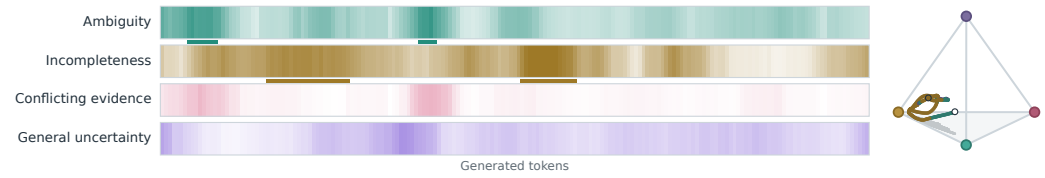


Figure 10: **Token-level readout and simplex, incompleteness.** Prompt, response and per-category readout of Qwen3.5, centered and smoothed as described in the text. Marked spans are the top ten percent of readings.

The installation manual specifies 40 Nm for the bolts. The manufacturer's service bulletin, issued the same month, specifies 55 Nm. What torque should I use?

You should use the 55 Nm specification. Here is the reasoning behind this decision: 1. Hierarchy of Authority: In engineering and manufacturing, a Service Bulletin (or Technical Service Bulletin/TSB) issued after the original installation manual serves as an official update or correction. It supersedes the original documentation. 2. Timing: Since the bulletin was issued the same month as the manual (or shortly thereafter), it likely addresses a critical finding, a calculation error in the original manual, or a specific batch of components that requires a higher clamping force to ensure safety or prevent failure. 3. Risk Management: Using the lower torque (40 Nm) when a higher value (55 Nm) has been explicitly mandated by the manufacturer could lead to joint failure, loosening under vibration, or safety hazards. Conversely, using the updated higher value is the only way to remain compliant with the manufacturer's current safety standards. Recommendation: Apply 55 Nm to the bolts. Additionally, it is good practice to document this decision in your work log or inspection report, noting that you followed the specific Service Bulletin (cite the bulletin number and date) rather than the original manual, to provide a clear audit trail for future reference.

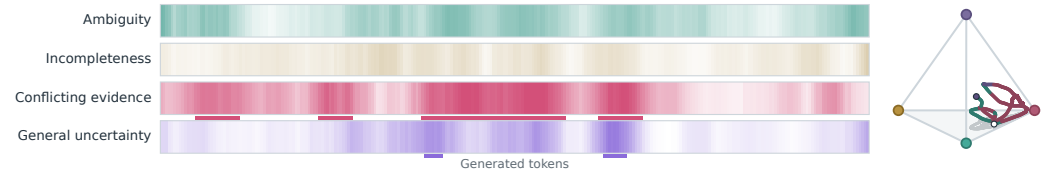


Figure 11: **Token-level readout and simplex, conflicting evidence.** Prompt, response and per-category readout of Qwen3.5, centered and smoothed as described in the text. Marked spans are the top ten percent of readings.

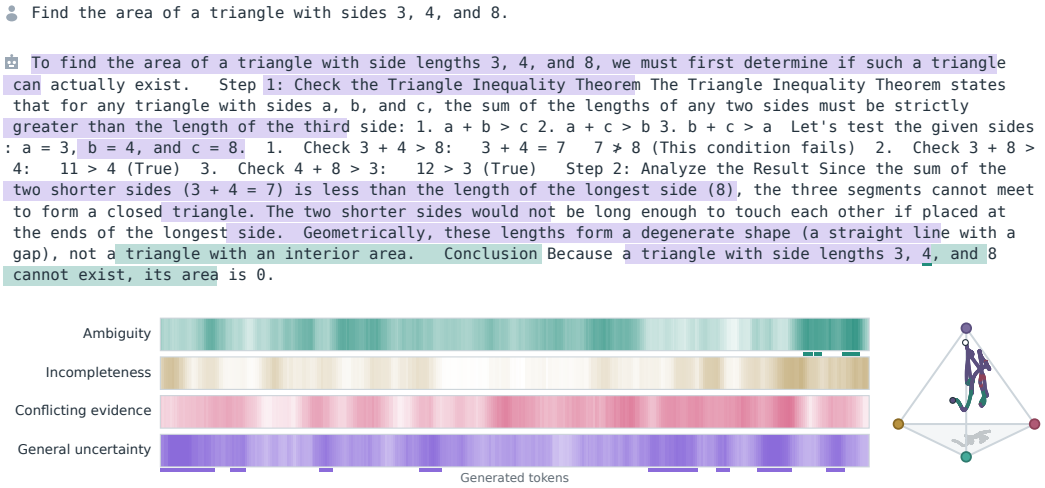


Figure 12: **Token-level readout and simplex, general uncertainty.** Prompt, response and per-category readout of Qwen3.5, centered and smoothed as described in the text. Marked spans are the top ten percent of readings.

so the trace before the intervention is identical across arms. Figures 13 to 16 show how the model reacts to increasing steering strength.

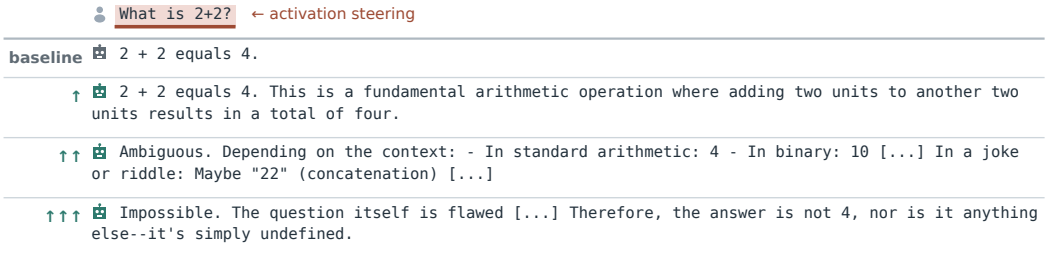


Figure 13: **Steering strength, ambiguity.** The prompt, the unsteered answer, and the continuation after adding the **ambiguity** direction to the prompt states at $\lambda = 0.04$, 0.08, and 0.09.

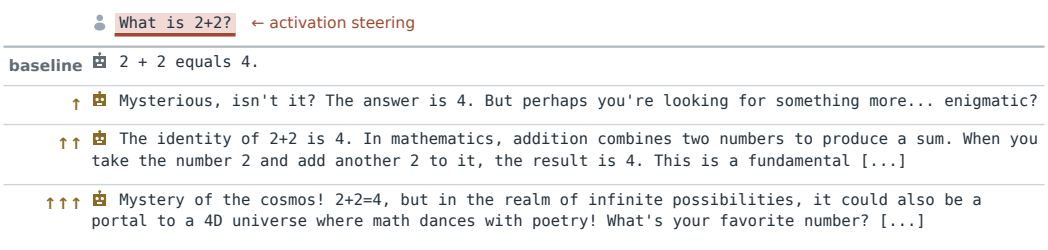


Figure 14: **Steering strength, incompleteness.** The prompt, the unsteered answer, and the continuation after adding the **incompleteness** direction to the prompt states at $\lambda = 0.04$, 0.08, and 0.09.

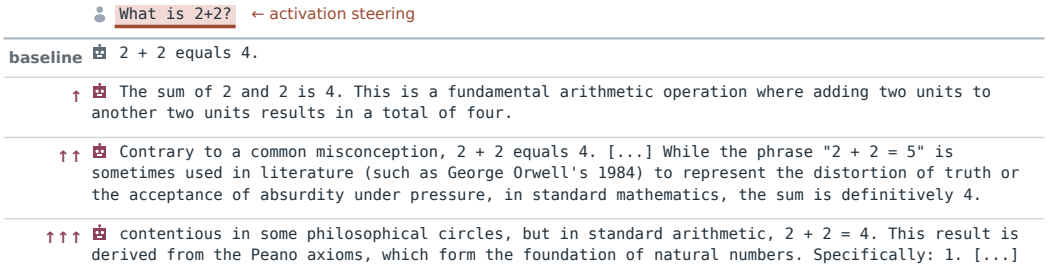


Figure 15: **Steering strength, conflicting evidence.** The prompt, the unsteered answer, and the continuation after adding the **conflicting evidence** direction to the prompt states at $\lambda = 0.04, 0.06,$ and 0.12.

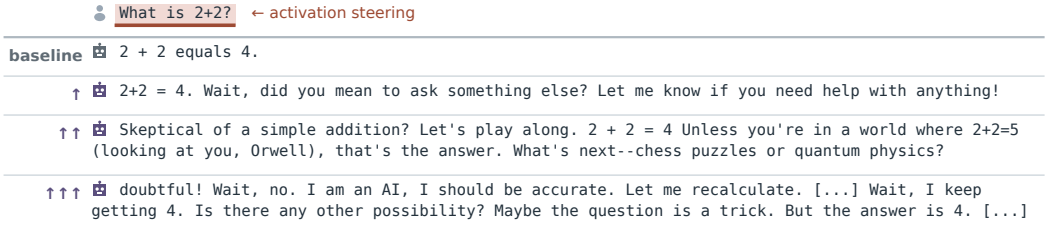


Figure 16: **Steering strength, general uncertainty.** The prompt, the unsteered answer, and the continuation after adding the **general uncertainty** direction to the prompt states at $\lambda = 0.03, 0.10,$ and 0.17.