TRAINING-FREE AND GENERATION-FREE ANSWERABILITY ESTIMATION IN FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models, including large language models (LLMs) and vision-language models (VLMs), answer some queries correctly and fail on others. We study this through the concept of answerability, i.e., whether a model can correctly answer a given query. Estimating answerability before generation enables adaptive inference, such as model routing, retrieval, or abstention. Most existing methods either estimate answerability after generation or rely on training data to construct an estimator. We consider a training-free and generation-free setting, in which the estimate requires neither parameter optimisation nor answer generation. We propose that answerability can be understood through three factors: input adequacy, knowledge boundary, and reasoning capability. This decomposition provides a structured view of why a query may or may not be answerable. Building on this decomposition, we propose Answerability Direction Probing (ADP), which constructs answerability-related concept directions from vocabulary tokens in the unembedding space and probes representations across layers, using a Jacobian lens for intermediate-layer alignment. The score is obtained from a single forward pass. With a single configuration shared across LLMs and VLMs, ADP outperforms other compared methods across multiple model families and diverse tasks. Beyond answerability estimation, we showcase that ADP can be used for model cascading, reducing inference cost without compromising performance.

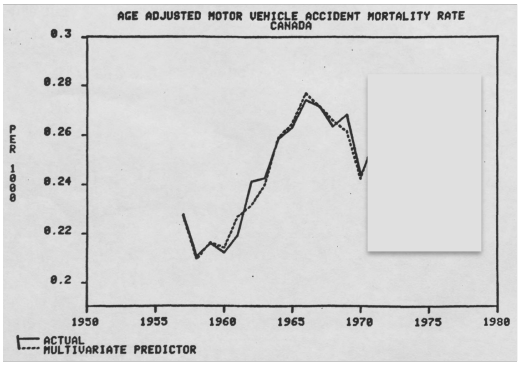
1 INTRODUCTION

Foundation models answer some queries correctly and fail on others. We study this through *answerability*: whether a model can correctly answer a given query. Estimating answerability before the model outputs the first token can avoid unnecessary generation cost, saving both computation and latency, especially when answers are long or several candidate answers need to be generated. This enables cheaper and faster adaptive inference: the system can retrieve additional information, route the query, or abstain before generating any token.

Many existing methods estimate answerability from generated answers (Kuhn et al., 2023; Farquhar et al., 2024). Other generation-free methods require task-specific training data (Kadavath et al., 2022; Cencerrado et al., 2026; Kogilathota et al., 2026). Internal Confidence (IC) (Chen et al., 2026) avoids both. It uses a self-assessment prompt and projects hidden states onto the unembedding vectors of the *Yes* and *No* tokens. This gives a training-free and generation-free answerability score for LLMs. However, IC reduces answerability to a fixed *Yes*/*No* signal. If the original query also has a *Yes*/*No* answer, the same logits may reflect both the answer itself and the model's answerability judgement. More generally, IC asks the model to compress multiple internal signals about answerability into a binary judgement, which may lose information.

We instead view answerability as depending on three factors. We decompose answerability into *Input Adequacy*, *Knowledge Boundary*, and *Reasoning Capability*. These factors describe whether the query gives enough information, whether the model has the needed knowledge, and whether it can reason over the available information. Figure 1 gives an example of each. Such signals may already be present in intermediate hidden states before the model maps them to a final *Yes*/*No* judgement (Gurnee et al., 2026). This motivates us to probe these signals directly from model representations.

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.



Q: What is the 'actual' value per 1000, during the year 1975?
GT: 0.28 VLM: 0.26

Input Adequacy

Q: What is George Rankin's occupation?
GT: Politician
LLM: Mathematician

Knowledge Boundary

Q: Brandon sold 86 geckos last year. He sold twice that many the year before. How many geckos has Brandon sold in the last two years?
GT: 2 * 86 = 172
86 + 172 = 258
LLM: 172

Reasoning Capability

Figure 1: Examples of the three sources of answerability failure: Input Adequacy (left), Knowledge Boundary (top right) and Reasoning Capability (bottom right). GT: reference answer; red: the model's wrong answer.

Under the Linear Representation Hypothesis, high-level properties can be represented approximately as directions in the model's representation space (Park et al., 2024). We therefore assume that there is an underlying answerability direction in this space. Prior work on additive compositionality suggests that a complex concept direction can be approximated by combining directions corresponding to its constituent concepts (Mikolov et al., 2013; Gittens et al., 2017). These constituent directions can also be estimated from the unembedding vectors of contrastive words associated with each concept (Park et al., 2024). Together, these observations suggest a way to estimate the underlying answerability direction: decompose answerability into constituent concepts, construct directions for these concepts, and combine selected directions.

From the three-part decomposition, we define answerability-related concepts including Evidence Availability, Content Familiarity, Certainty, etc. We construct a direction for each concept from contrastive words, combine selected directions to obtain an estimated answerability direction, and project the model's hidden representations onto this direction to obtain an answerability score. We call this framework *Answerability Direction Probing* (ADP). ADP allows the richer answerability-related information retained in the hidden states to contribute directly to the score, thereby avoiding reliance on two model-controlled transformations: compressing this information into a holistic binary judgement and mapping that judgement onto the prescribed contrast tokens.

In practice, ADP probes intermediate and later layers and aggregates the resulting scores. For intermediate layers, we use Jacobian-lens alignment (Gurnee et al., 2026) to map their representations into the final-layer space before projection, allowing the same constructed answerability direction to be applied consistently across layers. All layer-wise alignment and projection are performed on the hidden states obtained from a single forward pass.

To the best of our knowledge, ADP is the first training-free and generation-free answerability estimation method for foundation models that requires no task data and applies to both VLMs and LLMs without modality-specific changes to the estimator. Beyond the estimator itself, we introduce a three-part decomposition of answerability into Input Adequacy, Knowledge Boundary, and Reasoning Capability, providing a structured account of why a query may be unanswerable. ADP is lightweight, requiring only a single forward pass through the model at inference to obtain its answerability score. Despite the low cost, ADP substantially outperforms existing training-free and generation-free baselines overall across 12 models and 5 datasets. Further controlled comparisons show that the constructed answerability direction itself contributes substantially to these gains. Its pre-generation score also provides a lightweight signal for adaptive inference, as demonstrated through model cascading.

2 Related Work

Answerability estimation. Methods relevant to answerability estimation operate either after or before generation. Post-generation approaches first obtain one or more answers and then assess them, for example through semantic consistency across samples (Kuhn et al., 2023; Farquhar et al., 2024), by asking the model whether a generated answer is correct as in P(True) (Kadavath et al., 2022), or by probing representations associated with generated answers (Orgad et al., 2025). Pre-generation approaches instead estimate answerability before any output token is generated. P(IK) (Kadavath et al., 2022) fine-tunes the model to predict whether it will answer correctly, while Cencerrado et al. (2026) construct a correctness direction from labelled representations without parameter optimisation. Internal Confidence (IC) (Chen et al., 2026) requires neither parameter optimisation nor task samples in its zero-shot setting and estimates answerability through the model's preference for the *Yes*/*No* tokens under a self-assessment prompt. For VLMs, HALP (Kogilathota et al., 2026) shows that pre-generation internal representations contain useful answerability signals, but trains MLP probes on labelled task data. ADP instead targets training-free and generation-free answerability estimation without task data across both LLMs and VLMs.

Representation probing. Internal-state estimators also differ in how their probing signals are constructed. Data-derived approaches construct them from model representations, including labelled class differences and trained probes (Cencerrado et al., 2026; Orgad et al., 2025; Kogilathota et al., 2026); CCS (Burns et al., 2023) avoids labels but still learns a linear probe from contrast pairs drawn from task samples. Vocabulary-specified approaches instead define what is probed directly through selected tokens: IC uses the *Yes*/*No* contrast (Chen et al., 2026), while PROJECTAWAY (Jiang et al., 2025) uses target-object tokens to probe whether an object is represented. Under the Linear Representation Hypothesis, high-level properties can be represented approximately as directions in representation space (Park et al., 2024); related work has constructed such directions from opposing word groups (An et al., 2018). Applying vocabulary-specified probes to intermediate layers additionally raises the problem of representational drift. The logit lens (nostalgebraist, 2020) applies the final vocabulary projection directly, whereas the tuned lens (Belrose et al., 2023) and Jacobian lens (Gurnee et al., 2026) align intermediate representations before vocabulary projection. ADP follows the vocabulary-specified approach while constructing answerability-related directions without task samples and using Jacobian-lens alignment for multi-layer probing.

3 Method

3.1 Problem formulation

Let F_θ be a pre-trained model with L Transformer layers, hidden dimension d, vocabulary $\mathcal{V}$, and unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}|\times d}$ with rows $\mathbf{w}_t$. For an LLM, the input is a question $x = q$; for a VLM, $x = (I, q)$ additionally contains an image I. The model is presented with $\tilde{x} = \Pi(x)$, where Π places x in a fixed self-assessment prompt that asks whether the model can provide a fully correct answer and instructs it to reply only *Yes* or *No*. One forward pass on $\tilde{x}$ gives hidden states $\mathbf{h}_n^{(l)}(\tilde{x}) \in \mathbb{R}^d$ at layer l and position n.

The goal is a score

$$s(x) = g\Big(\{\mathbf{h}_n^{(l)}(\tilde{x})\}_{1\le l\le L,\ 1\le n\le|\tilde{x}|}\Big) \tag{1}$$

that ranks answerable model–input pairs above unanswerable ones and satisfies two constraints. First, it is *generation-free*: $s(x)$ depends on no generated token and is fixed before the first output token. Second, it is *training-free*: constructing s requires no parameter optimisation and no task samples, labelled or unlabelled.

3.2 Overview

A query may be unanswerable for different reasons. Motivated by this distinction, we decompose answerability failure into three factors,

$$\mathbf{z}_\theta(x) = \big(z_{\mathrm{IA}}(x), z_{\mathrm{KB},\theta}(x), z_{\mathrm{RC},\theta}(x)\big). \tag{2}$$

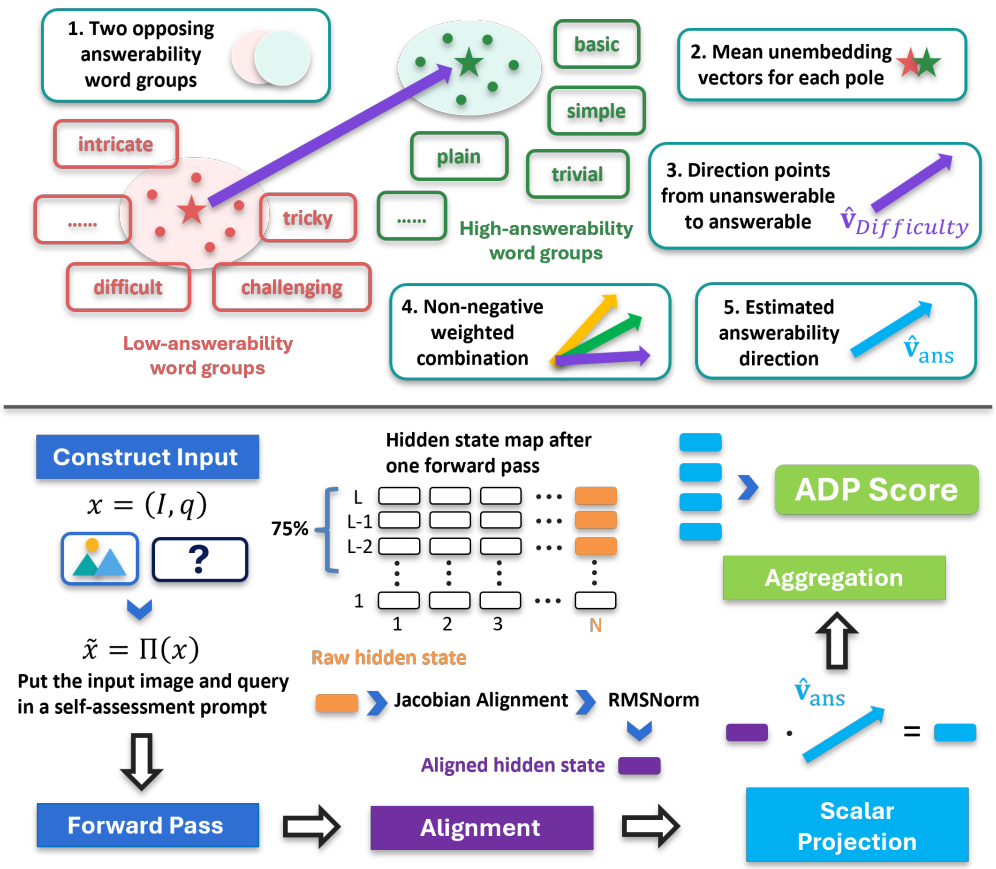


Figure 2: Overview of ADP. Top: each answerability-related direction is constructed as the difference between the mean unembedding vectors of two opposing seed-word groups. Selected directions are combined and normalised to produce the estimated answerability direction $\hat{\mathbf{v}}_{\text{ans}}$. Bottom: after one forward pass on the self-assessment prompt, hidden states at the last input position are taken from the probed layers. Intermediate-layer states are aligned to the final-layer representation space with the Jacobian lens, while the final-layer state requires no alignment. The aligned states are projected onto $\hat{\mathbf{v}}_{\text{ans}}$, and the resulting projections are averaged to obtain the ADP score.

Here, *Input Adequacy* ($z_{\text{IA}}(x)$) captures whether the input is sufficient and unambiguous for determining a correct answer given the required background knowledge. *Knowledge Boundary* ($z_{\text{KB},\theta}(x)$) captures whether the model has the required knowledge. *Reasoning Capability* ($z_{\text{RC},\theta}(x)$) captures whether the model can reason over the available information. Input adequacy concerns the query itself, while knowledge boundary and reasoning capability depend on the model. This decomposition does not assume that the three factors are orthogonal. Instead, it provides a conceptual basis for identifying answerability-related signals in model representations.

Following the Linear Representation Hypothesis (Park et al., 2024), ADP assumes an underlying answerability direction $\mathbf{v}^* \in \mathbb{R}^d$ in the final-layer representation space, along which larger projections indicate greater support for answerability. Because $\mathbf{v}^*$ is unknown, ADP estimates it from directions corresponding to answerability-related concepts derived from the decomposition. Motivated by additive compositionality (Mikolov et al., 2013; Gittens et al., 2017), selected concept directions are combined into a single estimated answerability direction (Section 3.3). ADP then applies this direction to representations from multiple intermediate and later layers after aligning them to the final-layer space (Section 3.4). Figure 2 summarises the procedure.

3.3 Direction Construction

For each answerability-related concept a, ADP specifies two opposing token sets $\mathcal{S}_a^H, \mathcal{S}_a^L \subseteq \mathcal{V}$, representing the unanswerable and answerable poles, respectively. Each pole is represented by the

mean of its unembedding vectors,

$$\boldsymbol{\mu}_a^H = \frac{1}{|\mathcal{S}_a^H|} \sum_{t \in \mathcal{S}_a^H} \mathbf{w}_t, \qquad \boldsymbol{\mu}_a^L = \frac{1}{|\mathcal{S}_a^L|} \sum_{t \in \mathcal{S}_a^L} \mathbf{w}_t, \tag{3}$$

and the corresponding concept direction is

$$\mathbf{v}_a = \boldsymbol{\mu}_a^L - \boldsymbol{\mu}_a^H. \tag{4}$$

For a representation after the model's final normalisation, projection onto $\mathbf{v}_a$ equals the difference between the mean unembedding scores of the answerable-side and unanswerable-side tokens. The individual concept directions are not normalised before combination.

The seed words are specified manually in advance to represent the intended opposing concepts with as little ambiguity as possible. Each seed word must form a single vocabulary token when preceded by a space under the target tokenizer, so that it corresponds to one row of the unembedding matrix. No target-task samples are used to construct the directions. The concrete seed sets and tokenisation details are given in Appendix A.

We organise the answerability-related concepts at two levels. Five *factor-level* directions capture more specific concepts associated with the three-part decomposition: Clarity, Evidence, and Ambiguity for Input Adequacy; Familiarity for Knowledge Boundary; and Certainty for Reasoning Capability. Two *integrative* directions capture broader judgements: Challenge, which reflects the overall demand of the query, and Feasibility, which reflects how feasible a correct answer is for the current model–input pair.

To estimate the underlying answerability direction $\mathbf{v}^*$, ADP combines a selected set $\mathcal{A}$ of concept directions using non-negative weights,

$$\mathbf{v}_{\text{ans}} = \sum_{a \in \mathcal{A}} \lambda_a \mathbf{v}_a, \qquad \lambda_a \geq 0, \tag{5}$$

which preserves their shared polarity. Assuming $\mathbf{v}_{\text{ans}} \neq \mathbf{0}$, the combined direction is then normalised,

$$\hat{\mathbf{v}}_{\text{ans}} = \frac{\mathbf{v}_{\text{ans}}}{\|\mathbf{v}_{\text{ans}}\|}. \tag{6}$$

The selected directions and their weights are fixed without target-task samples. The configuration used in the main experiments is specified in Section 4.

3.4 Aggregation

Let $N = |\tilde{x}|$ denote the last input position. Under causal attention, its representation has access to the complete input sequence. Let $\mathcal{L} \subseteq \{1, \ldots, L\}$ be the probed layers, chosen as a fixed span of intermediate and later layers ending at the final layer.

The concept directions are defined in the final-layer representation space. For an intermediate layer l, ADP therefore uses the Jacobian lens (Gurnee et al., 2026) to approximate its downstream mapping into this space. Let $\mathbf{J}_l \in \mathbb{R}^{d \times d}$ denote the averaged Jacobian from layer l to the final layer, estimated on task-independent text and then fixed. The final layer requires no mapping, so $\mathbf{J}_L = \mathbf{I}$.

After alignment, ADP applies the model's final normalisation:

$$\mathbf{h}_{N,J}^{(l)}(\tilde{x}) = \mathrm{Norm}_f\left(\mathbf{J}_l \mathbf{h}_N^{(l)}(\tilde{x})\right), \tag{7}$$

where Norm_f denotes the model's final normalisation module. The aligned representation is then projected onto the estimated answerability direction,

$$r_N^{(l)}(\tilde{x}) = \hat{\mathbf{v}}_{\text{ans}}^\top \mathbf{h}_{N,J}^{(l)}(\tilde{x}). \tag{8}$$

Because $\hat{\mathbf{v}}_{\text{ans}}$ points from the unanswerable pole towards the answerable pole, larger values indicate stronger support for answerability.

The final ADP score is the unweighted mean of the layer-wise projections,

$$s(x) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} r_N^{(l)}(\tilde{x}). \tag{9}$$

All layer-wise scores are computed from hidden states obtained in the same forward pass on $\tilde{x}$, while the Jacobians are computed offline on task-independent text. Once the Jacobians are fixed, each non-final probed layer adds one $d \times d$ matrix–vector product; every probed layer additionally requires the final normalisation and one inner product.

4 EXPERIMENTAL SETUP

Models and tasks. We evaluate 12 models across 5 datasets: six VLMs, Qwen3-VL 2B, 4B, 8B and 32B (Qwen Team, 2025) and Gemma-3 4B and 12B (Gemma Team, 2025), on DocVQA (Mathew et al., 2021), OK-VQA (Marino et al., 2019) and MathVista (Lu et al., 2024); and six LLMs, Llama-3.2 1B and 3B (Grattafiori et al., 2024), Phi-4-mini and Phi-4 (Abdin et al., 2024), and Qwen3 0.6B and 8B (Yang et al., 2025), on PopQA (Mallen et al., 2023) and GSM8K (Cobbe et al., 2021). Together, these tasks provide broad coverage of the three sources of answerability failure introduced in Section 3.2. We use the DocVQA validation set, OK-VQA val2014, and the MathVista and GSM8K test sets. For PopQA, we take the first 200 examples from each of its 16 relation types and remove duplicates, leaving 3,021 questions. This gives 30 model–task combinations.

Labels. Each model answers every question by greedy decoding. MathVista and GSM8K use chain-of-thought prompting, while the other tasks use short-answer prompting. For the three VQA datasets, Gemini 3.7 Flash judges the model answers using the image, question and reference answer. It first checks whether the input is sufficient to determine a correct answer, then validates the reference as needed and grades the model answers; all model answers to questions with insufficient input are labelled incorrect. PopQA uses normalised whole-word matching against accepted aliases, and GSM8K compares the parsed final number with the reference. A model–input pair is labelled answerable if its answer is correct and unanswerable otherwise. Full prompting and judging details are given in Appendix B.

Baselines and reference. We compare ADP with six training-free and generation-free baselines and include ADP(Yes) as a controlled reference. Following the raw-query protocol of Chen et al. (2026), four likelihood-based scores are computed over the question tokens without the self-assessment prompt: Max-logp (Manakul et al., 2023), predictive entropy (PE) (Malinin & Gales, 2021), Min-K (Shi et al., 2024) and perplexity (PPL). We negate these scores so that larger values indicate greater confidence. P(Yes) applies a two-way softmax to the *Yes*/*No* logits at the final layer and last input position. Internal Confidence (IC) (Chen et al., 2026) aggregates the corresponding P(Yes) values over the last six positions and the same probed layer span as ADP using its attenuating weights. ADP(Yes) keeps the complete ADP pipeline but replaces $\hat{\mathbf{v}}_{\mathrm{ans}}$ with a unit *Yes*/*No* direction. ADP and ADP(Yes) therefore differ only in the probing direction. Exact token forms are given in Appendix B.

ADP configuration. We use a single ADP configuration across all models and tasks: $\mathcal{A} = \{\mathrm{Challenge}, \mathrm{Feasibility}\}$ with equal weights $\lambda_a = 1$, the last 75% of layers including the final layer, and the last input position. We use the two integrative directions because they express broader answerability judgements rather than targeting any single failure source. The five factor-level directions are used only for analysis. Ablations of these choices are reported in Section 5.3.

Implementation. The fixed self-assessment prompt is wrapped in each model's chat template. The Jacobians are computed by automatic differentiation and averaged over 400 task-independent WikiText-103 passages (Merity et al., 2017), without updating model parameters. Exact prompts, model-specific template settings, precision and tokenisation details are given in Appendix B.

Metrics. We report AUROC and prediction rejection ratio (PRR) (Malinin et al., 2017), both multiplied by 100. AUROC measures how well a score ranks answerable above unanswerable model–input pairs. PRR measures selective rejection quality when the lowest-scored inputs are rejected, with 100 corresponding to oracle rejection and 0 to random rejection; negative values indicate worse-than-random rejection. Both metrics are computed separately for each model–task combination and then macro-averaged.

Table 1: AUROC (×100) of the eight scores on the 30 model–task combinations, grouped by model family (VLMs above the rule, LLMs below). Best per row in bold, second best underlined.

Model	Task	Max-logp	PE	Min-K	PPL	P(Yes)	IC	ADP(Yes)	ADP
Qwen3-VL									
Qwen3-VL-2B	DocVQA	56.70	61.12	59.06	55.23	71.93	72.67	74.46	**76.26**
	OK-VQA	58.58	60.25	60.81	60.35	73.41	72.84	72.94	**73.63**
	MathVista	45.12	52.13	48.81	49.04	60.05	62.62	62.32	**68.52**
Qwen3-VL-4B	DocVQA	55.89	56.81	56.67	58.75	64.76	69.00	74.14	**81.72**
	OK-VQA	56.87	55.95	56.20	57.07	74.52	74.54	75.04	**75.90**
	MathVista	43.68	49.06	47.98	51.01	62.13	59.30	65.47	**69.61**
Qwen3-VL-8B	DocVQA	57.07	64.17	63.30	58.80	66.72	73.01	74.29	**82.69**
	OK-VQA	55.82	54.54	55.09	56.18	75.34	74.29	75.74	**76.11**
	MathVista	44.54	46.72	45.81	49.71	66.90	65.53	69.83	**73.57**
Qwen3-VL-32B	DocVQA	60.91	59.81	59.60	58.18	67.39	73.13	75.62	**82.23**
	OK-VQA	53.83	51.10	51.77	53.63	80.00	79.68	80.58	**81.21**
	MathVista	43.07	46.95	46.72	51.84	64.46	68.92	70.18	**76.72**
Gemma-3									
Gemma-3-4B	DocVQA	50.06	52.22	52.23	52.94	52.03	50.58	**64.03**	63.64
	OK-VQA	54.76	55.50	55.58	54.53	65.69	66.14	70.04	**70.06**
	MathVista	49.75	53.93	53.19	52.37	57.98	60.07	63.93	**64.45**
Gemma-3-12B	DocVQA	53.38	58.45	58.12	56.39	53.63	64.29	67.24	**71.20**
	OK-VQA	50.40	50.70	50.80	52.08	70.22	71.84	71.82	**72.55**
	MathVista	51.24	50.91	51.76	51.46	58.35	62.05	64.55	**67.39**
Llama-3.2									
Llama-3.2-1B	PopQA	53.00	56.62	58.10	56.49	45.75	47.84	71.98	**78.99**
	GSM8K	53.92	**63.48**	56.70	50.88	41.04	41.09	58.19	52.12
Llama-3.2-3B	PopQA	59.35	64.35	68.59	67.88	78.65	78.49	75.50	**79.27**
	GSM8K	53.16	**64.93**	56.26	51.76	55.05	54.61	52.87	61.02
Phi-4									
Phi-4-mini	PopQA	46.79	53.08	53.85	53.35	69.92	69.94	69.21	**72.83**
	GSM8K	53.29	64.79	58.06	58.45	66.43	66.43	64.86	**68.00**
Phi-4	PopQA	51.16	59.23	59.61	61.58	74.71	77.51	76.40	**78.22**
	GSM8K	60.74	61.24	58.77	60.72	56.08	70.28	71.61	**74.28**
Qwen3									
Qwen3-0.6B	PopQA	68.26	52.83	61.12	60.10	52.70	52.98	68.22	**69.91**
	GSM8K	50.84	**68.42**	58.64	56.90	54.29	53.81	56.16	61.33
Qwen3-8B	PopQA	55.20	58.33	59.71	59.86	61.29	58.34	70.34	**76.02**
	GSM8K	51.78	59.59	54.78	52.07	51.01	46.14	51.44	**59.81**

Table 2: Macro-average AUROC and PRR (×100) over the 18 VLM and 12 LLM model–task combinations, together with the overall macro-average over all 30 combinations. Best per row in bold, second best underlined.

Metric	Models	Max-logp	PE	Min-K	PPL	P(Yes)	IC	ADP(Yes)	ADP
AUROC	VLM	52.32	54.46	54.08	54.42	65.86	67.81	70.68	**73.75**
	LLM	54.79	60.57	58.68	57.50	58.91	59.79	65.56	**69.32**
	Overall	53.31	56.91	55.92	55.65	63.08	64.60	68.63	**71.98**
PRR	VLM	7.15	10.87	10.59	11.87	34.42	42.08	48.96	**56.91**
	LLM	12.39	29.53	24.06	19.03	18.13	20.75	32.04	**40.34**
	Overall	9.24	18.33	15.98	14.73	27.90	33.54	42.19	**50.28**

5 RESULTS

5.1 MAIN RESULTS

Table 1 reports AUROC for all 30 model–task combinations, while Table 2 gives the corresponding macro averages. ADP achieves the highest AUROC in 26 of the 30 combinations and the highest overall macro-average AUROC, at 71.98. It is also best when VLMs and LLMs are considered separately, reaching 73.75 and 69.32, respectively. The same pattern holds for PRR, where ADP reaches an overall macro-average of 50.28.

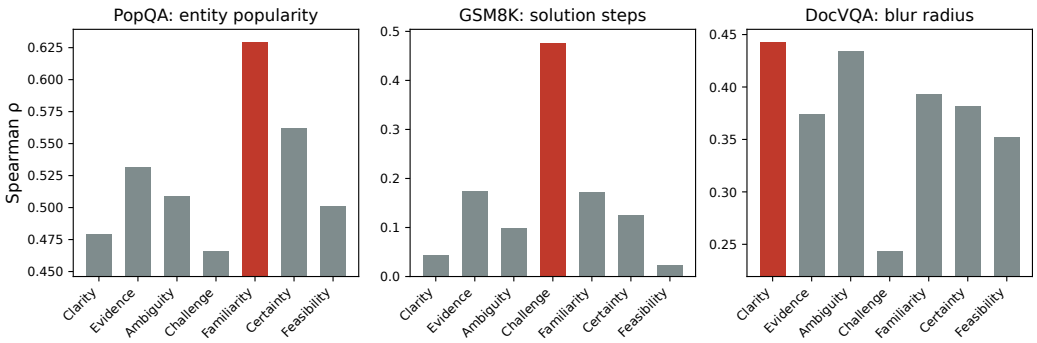


Figure 3: Spearman correlation between each direction's projection and a factor-relevant variable. Left: PopQA entity rarity. Middle: GSM8K reference-solution steps. Right: Gaussian blur applied to DocVQA images. Results are averaged over the selected models; per-model results and selection details are given in Appendix D.

The comparison with ADP(Yes) isolates the contribution of the constructed answerability direction. ADP and ADP(Yes) share the same prompt, probing position, layer span, Jacobian alignment, normalisation and aggregation, differing only in the probing direction; replacing the *Yes*/*No* direction with the constructed answerability direction raises macro-average AUROC from 70.68 to 73.75 on VLMs and from 65.56 to 69.32 on LLMs. ADP(Yes) also improves substantially over P(Yes), although this difference reflects several changes in the probing pipeline and is therefore not attributed to a single component. The four AUROC exceptions are Gemma-3-4B on DocVQA, where ADP(Yes) is 0.39 points higher, and GSM8K on Qwen3-0.6B, Llama-3.2-1B and Llama-3.2-3B, where PE performs best; ADP is best on GSM8K for the other three LLMs. Two observations help characterise this pattern. PE sums the negative log-likelihood over the question tokens and is therefore length-dependent; PPL, which normalises the same quantity by token count, drops from 63.74 to 55.13 in macro-average AUROC, suggesting that sequence length contributes substantially to PE's signal on GSM8K. For the same three models, P(Yes) on GSM8K lies between 41 and 55 AUROC, indicating that the explicit *Yes*/*No* self-assessment signal is weak in these combinations. The PRR results (Appendix C) show the same pattern: ADP is best in all 18 VLM combinations and has the highest LLM macro-average PRR, while PE remains strongest on GSM8K.

5.2 What the directions measure

We examine whether the seven directions respond to variables associated with their intended concepts. For this analysis, we reverse the projection sign so that larger values correspond to the unanswerable pole. We use entity rarity on PopQA, the number of lines of the reference solution as a proxy for reasoning steps on GSM8K, and controlled image blur on DocVQA; full details are given in Appendix D.

Figure 3 shows selective but overlapping responses. Familiarity responds most strongly to entity rarity on PopQA, while Challenge responds most strongly to the number of reasoning steps on GSM8K. Blur produces a broader response, with Clarity and Ambiguity among the strongest and Challenge among the weakest. These patterns support the intended interpretations of the directions without implying that they form independent or factor-exclusive coordinates.

5.3 Ablation and robustness

On VLMs, replacing the last input position by the best alternative position reduces macro-average AUROC by 4.31 points, while using Challenge or Feasibility alone reduces it by 1.58 and 2.69 points, respectively. Probing only the final layer reduces AUROC by 1.00 point. Analyses of seed-word robustness, direction composition, and layers and positions are given in Appendices F, G and H.

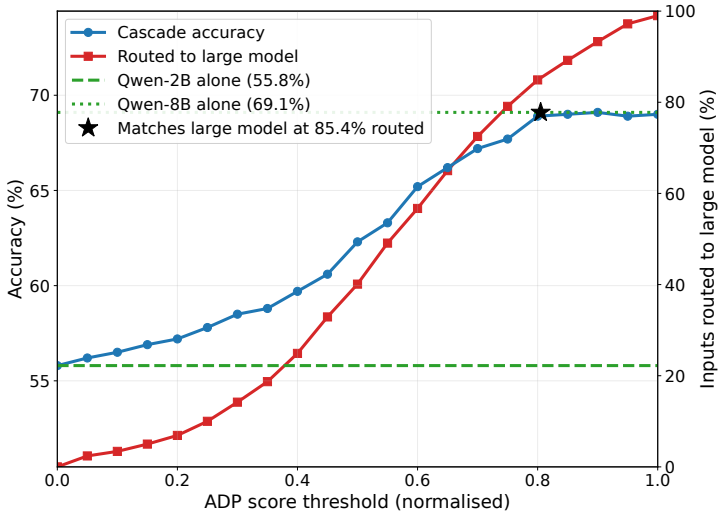


Figure 4: Cascading Qwen3-VL-2B to Qwen3-VL-8B on MathVista using the small model's ADP score. At the operating point matching the large model's accuracy (69.1%), 85.4% of inputs are routed to it.

5.4 Model Cascading

Pre-generation answerability estimation can support model cascading because the routing decision is made before the small model generates an answer. On MathVista, we cascade Qwen3-VL-2B (55.8% accuracy) to Qwen3-VL-8B (69.1%) by routing inputs with lower small-model ADP scores to the larger model. At the operating point that matches the large model's 69.1% accuracy, the cascade routes 85.4% of inputs to it. Thus, in this case study, ADP reduces the fraction of inputs sent to the larger model by 14.6 percentage points while matching its accuracy, without first generating an answer from the smaller model (Figure 4).

6 Conclusion

We presented Answerability Direction Probing (ADP), which estimates whether a foundation model can answer a query from a single forward pass, before any answer token is generated. Guided by a three-part decomposition of answerability failure, ADP constructs answerability-related directions from vocabulary tokens, combines selected directions into an estimated answerability direction, and probes Jacobian-aligned hidden states across layers. With one fixed configuration, ADP achieves the highest AUROC among the compared training-free and generation-free scores on 26 of 30 model–task combinations across 12 models spanning LLMs and VLMs. The controlled comparison with the same pipeline using the *Yes*/*No* direction shows that the constructed direction contributes substantially to this improvement, while the individual directions respond to factor-relevant variables in ways consistent with their intended interpretations. The resulting pre-generation score also supports model cascading without first generating an answer from the smaller model.

Limitations. ADP relies on answerability-related information represented during the model's self-assessment computation, and this signal can be weak. On GSM8K, for example, the explicit *Yes*/*No* self-assessment signal is weak for several smaller models, and a length-sensitive token-likelihood score performs better. The two integrative directions are a fixed design choice rather than a uniquely determined decomposition; post-hoc analyses support this choice, but other seed words or direction sets meeting the same requirements could instantiate the method, and we have not searched that space.

AI USE STATEMENT

Generative AI tools were used to assist with the implementation, review and debugging of analysis and plotting scripts, and with language and consistency review of text written by the authors. All experiments were designed, configured and run by the authors, and all experimental results and claims were independently verified by the authors. No content was incorporated without the authors' own verification and revision. We take responsibility for the final content of this work.

REFERENCES

Marah Abdin et al. Phi-4 technical report. *arXiv preprint*, arXiv:2412.08905, 2024.

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2450–2461, 2018.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint*, arXiv:2303.08112, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting LLM answer accuracy from question-only linear probes. In *ICLR 2026 Workshop on Predictable and Reliable AI*, 2026. URL https://openreview.net/forum?id=1QcY6LPcdQ.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models. In *International Conference on Learning Representations*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint*, arXiv:2110.14168, 2021.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint*, arXiv:2503.19786, 2025.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. Skip-gram $-$ zipf $+$ uniform $=$ vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 69–76, 2017.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783, 2024.

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. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, 2025.

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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom

Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint*, arXiv:2207.05221, 2022.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.287.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 45–50, 2017.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

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, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 746–751, 2013.

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

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025.

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, pp. 39643–39666, 2024.

Qwen Team. Qwen3-VL technical report. *arXiv preprint*, arXiv:2511.21631, 2025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint*, arXiv:2505.09388, 2025.

A Directions and seed words

Table 3 summarises the seven answerability-related directions with two seed words per pole.

Table 3: Answerability-related directions with two example seed words per pole; the full seed sets are in Table 4.

Level	Direction	Unanswerable pole	Answerable pole
Factor (input adequacy)	Clarity	blurry, fuzzy	sharp, crisp
Factor (input adequacy)	Evidence	absent, hidden	present, visible
Factor (input adequacy)	Ambiguity	unclear, vague	clear, precise
Factor (knowledge boundary)	Familiarity	unknown, unfamiliar	known, familiar
Factor (reasoning capability)	Certainty	unsure, uncertain	sure, certain
Integrative	Challenge	difficult, challenging	basic, simple
Integrative	Feasibility	cannot, unable	able, capable

ADP(Yes) uses the tokens " yes" and " no" (lower case, leading space); P(Yes) and IC use "Yes" and "No". Table 4 lists the 25 directions used in the analyses: the seven answerability-related directions, six control directions (polar and paired, unrelated to answerability), five related directions (paired, non-polar) and five random directions (no polarity or pairing); the two reference directions Yes/No and positive/negative are listed for completeness.

Table 4: Seed words used for the answerability-related, control, related and random directions.

Direction class	**Direction**	**Seed words**
Answerability-related	Clarity	*Pole A:* blurry, fuzzy, faint, faded
		Pole B: sharp, crisp, distinct, vivid
	Evidence	*Pole A:* absent, hidden, missing, obscured
		Pole B: present, visible, shown, displayed
	Ambiguity	*Pole A:* unclear, vague, ambiguous, confusing
		Pole B: clear, precise, specific, explicit
	Challenge	*Pole A:* difficult, challenging, tricky, intricate
		Pole B: basic, simple, plain, trivial
	Familiarity	*Pole A:* unknown, unfamiliar, unrecognized, rare
		Pole B: known, familiar, recognized, common
	Certainty	*Pole A:* unsure, uncertain, hesitant, doubtful
		Pole B: sure, certain, confident, convinced
	Feasibility	*Pole A:* cannot, unable, incapable, impossible
		Pole B: able, capable, possible, feasible

Continued on next page

Table 4 – continued from previous page

Direction class	Direction	Seed words
Control	Moisture	*Pole A:* wet, damp, moist, soaked
		Pole B: dry, dried, dusty, barren
	Loudness	*Pole A:* loud, noisy, booming, piercing
		Pole B: quiet, silent, muted, soothing
	Price	*Pole A:* expensive, costly, luxurious, premium
		Pole B: cheap, affordable, modest, bargain
	Softness	*Pole A:* soft, tender, fluffy, silky
		Pole B: firm, solid, sturdy, rigid
	Heaviness	*Pole A:* heavy, hefty, bulky, massive
		Pole B: light, airy, slim, delicate
	Quickness	*Pole A:* swift, rapid, hurried, fleeting
		Pole B: slow, sluggish, gradual, lingering
Related	Country–Capital	*Pole A:* France, Japan, Italy, Spain
		Pole B: Paris, Tokyo, Rome, Madrid
	Country–Language	*Pole A:* Germany, Poland, Greece, Turkey
		Pole B: German, Polish, Greek, Turkish
	Whole–Part	*Pole A:* tree, car, book, roof
		Pole B: branch, wheel, page, tile
	Profession–Tool	*Pole A:* chef, painter, doctor, tailor
		Pole B: knife, brush, needle, thread
	Source–Product	*Pole A:* bee, hen, maple, olive
		Pole B: honey, egg, syrup, oil
Random	Random 1	*Pole A:* bridge, cabbage, carriage, banner
		Pole B: button, tiger, gravel, thunder
	Random 2	*Pole A:* pencil, silver, desert, butter
		Pole B: spoon, blanket, cart, column
	Random 3	*Pole A:* mirror, wrench, kettle, shelf
		Pole B: candle, feather, letter, canyon
	Random 4	*Pole A:* harbor, bench, pigeon, almond
		Pole B: carpet, tunnel, socket, harness
	Random 5	*Pole A:* fork, canvas, lantern, island
		Pole B: helmet, onion, chair, window

B EXPERIMENTAL DETAILS

Answer generation. Each model answers each question by greedy decoding with at most 2048 new tokens. DocVQA, OK-VQA and PopQA use the short-answer prompt

```
Answer the following question with the shortest possible
answer (a word, number, or phrase) without explanation.
Question: {question} Answer:
```

and MathVista and GSM8K the chain-of-thought prompt

```
Answer the following question. First, give a brief
step-by-step reasoning. Keep each step concise and no more
than one short sentence. Then, on the last line, write
exactly 'Answer: <answer>', where <answer> is the shortest
```

```
possible answer (a word, number, or phrase).  Question:
{question}
```

from whose last `Answer:` line the answer is parsed.

Judge. VQA answers are judged by Gemini 3.7 Flash from the image, the question, the reference answer and all anonymised candidate answers for that question. The judge first decides whether a fully correct answer can be determined or verified from the given input together with the world knowledge the task requires; questions that fail this check are marked aleatoric and all their answers are counted wrong (6.1% of OK-VQA, 1.2% of MathVista, 0.1% of DocVQA). It then validates the reference, correcting it where necessary, and grades each candidate. One OK-VQA question was dropped because the judge refused the image. PopQA answers are correct if, after normalisation, they contain any accepted alias as a whole word; GSM8K answers are correct if the last number in the parsed answer equals the reference.

Self-assessment prompt. The user message is the image (VLMs only) followed by the text below, with the question substituted for `{question}`; the chat template adds the surrounding special tokens and the assistant prefix, and the probing position is the last token of the rendered input.

```
Original question:  {question}
Your task is to judge whether you can provide a fully
correct answer to the original question.
Reply with Yes if you can provide a fully correct answer to
the original question.
Reply with No if you cannot provide a fully correct answer
to the original question.
Judge:
```

For the position analyses on VLMs, 20 candidate positions are defined by block structure: the end-of-image token, the last three tokens of the question and task blocks, the Yes/No word and the last two tokens of each mapping line, the two cue tokens and the five tokens of the assistant prefix.

Jacobian lens and layers. $\mathbf{J}_l$ is estimated with the official implementation on the first 400 WikiText-103 texts, from each source layer in the last 75% of layers (excluding the final layer) to the final layer. The probed set $\mathcal{L}$ is the last 75% of layers including the final layer, rounded to the nearest layer: 21 of 28 (Qwen3-VL-2B, Qwen3-0.6B, Llama-3.2-3B), 27 of 36 (Qwen3-VL-4B, Qwen3-VL-8B, Qwen3-8B), 48 of 64 (Qwen3-VL-32B), 26 of 34 (Gemma-3-4B), 36 of 48 (Gemma-3-12B), 12 of 16 (Llama-3.2-1B), 24 of 32 (Phi-4-mini) and 30 of 40 (Phi-4).

LLM specifics. PopQA is evaluated on the first 200 questions of each of its 16 relation types (3 021 questions after removing duplicates). The Llama-3 chat template inserts the current date into its system header through `strftime_now`; we pass the template's default date so that hidden states do not depend on the run date. Qwen3 models are run with `enable_thinking=False`, which appends an empty think block to the assistant prefix; the probing position is the last token of that prefix.

Precision. Hidden states, Jacobians and unembedding rows are kept in bfloat16 and the projections are averaged in float64. Computing the projections in float32 instead changes no per-combination AUROC by more than 0.04 and no macro average at two decimals.

C PER-COMBINATION PRR

Table 5: PRR ($\times 100$) of the eight scores on the 30 model–task combinations. Negative values mean that rejecting by the score is worse than random rejection. Best per row in bold, second best underlined.

Model	Task	Max-logp	PE	Min-K	PPL	P(Yes)	IC	ADP(Yes)	ADP
Qwen3-VL									
Qwen3-VL-2B	DocVQA	9.86	28.30	23.63	13.19	47.27	55.28	56.34	**61.21**
	OK-VQA	21.25	23.91	24.53	23.14	54.06	53.77	53.50	**54.82**
	MathVista	-0.89	8.53	5.68	-1.26	26.31	38.03	34.40	**46.46**
Qwen3-VL-4B	DocVQA	20.19	20.66	20.56	26.75	30.07	42.93	52.09	**74.17**
	OK-VQA	18.63	16.60	16.83	18.04	56.06	55.61	59.89	**61.05**
	MathVista	-15.02	-2.91	-2.99	3.91	28.30	19.95	43.29	**52.24**
Qwen3-VL-8B	DocVQA	14.73	35.58	33.96	25.23	34.00	51.12	48.94	**74.07**
	OK-VQA	19.11	14.83	15.69	17.40	60.42	59.53	60.76	**62.59**
	MathVista	-12.04	-8.76	-8.87	1.01	37.94	40.99	49.87	**58.95**
Qwen3-VL-32B	DocVQA	29.52	20.42	20.13	24.23	35.49	58.18	54.17	**74.05**
	OK-VQA	14.44	7.31	8.46	12.84	68.01	68.35	69.81	**72.38**
	MathVista	-17.07	-12.16	-11.45	1.93	30.17	42.89	54.62	**65.82**
Gemma-3									
Gemma-3-4B	DocVQA	3.42	2.55	2.58	6.67	3.22	-0.83	31.84	**31.96**
	OK-VQA	11.62	14.46	14.56	10.48	30.64	33.89	48.21	**48.25**
	MathVista	-0.88	5.06	3.83	4.96	13.58	23.01	36.61	**36.89**
Gemma-3-12B	DocVQA	11.05	17.66	16.99	17.10	6.62	32.10	38.27	**45.73**
	OK-VQA	1.37	2.30	2.44	4.59	43.11	54.08	50.28	**57.79**
	MathVista	-0.61	1.33	4.15	3.47	14.23	28.52	38.33	**45.98**
Llama-3.2									
Llama-3.2-1B	PopQA	21.11	28.66	27.82	18.11	-10.90	-9.32	38.03	**48.75**
	GSM8K	4.75	**25.62**	12.84	3.09	-20.16	-18.33	11.83	3.13
Llama-3.2-3B	PopQA	25.79	40.82	44.34	37.41	**59.89**	58.95	54.14	59.03
	GSM8K	7.72	**38.10**	18.06	7.89	6.89	5.81	2.29	23.08
Phi-4									
Phi-4-mini	PopQA	3.66	25.21	24.07	16.30	46.97	46.23	45.72	**52.50**
	GSM8K	10.49	36.06	21.65	24.74	**41.99**	40.74	36.83	40.69
Phi-4	PopQA	9.20	33.09	32.51	27.06	46.93	**59.52**	56.07	58.75
	GSM8K	25.88	32.29	29.50	30.46	12.15	46.58	49.08	**59.12**
Qwen3									
Qwen3-0.6B	PopQA	17.86	3.13	12.32	16.69	8.84	9.57	32.04	**34.77**
	GSM8K	2.41	**34.94**	18.92	15.20	7.62	6.68	11.83	22.62
Qwen3-8B	PopQA	11.82	30.36	30.68	25.04	15.65	10.95	44.97	**57.24**
	GSM8K	7.92	**26.03**	15.97	6.33	1.65	-8.44	1.60	24.36

On VLMs ADP has the best PRR in all 18 combinations; on LLMs PE is ahead on four of the six GSM8K combinations (the three models of at most 3B parameters discussed in Section 5.1, and Qwen3-8B by 1.7 points), while ADP leads on PopQA and overall.

Bootstrap. A paired bootstrap over questions within each VLM combination ($B = 2000$; the same resampled indices for all scores; AUROC recomputed per combination and macro-averaged) gives 95% confidence intervals for the macro-average AUROC difference between ADP and each score: Max-logp $+21.43\,[+20.52, +22.37]$, PE $+19.28\,[+18.39, +20.16]$, Min-K $+19.66\,[+18.79, +20.52]$, PPL $+19.33\,[+18.47, +20.18]$, P(Yes) $+7.89\,[+7.44, +8.32]$, IC $+5.94\,[+5.44, +6.42]$, ADP(Yes) $+3.06\,[+2.69, +3.44]$. All seven intervals exclude zero.

D FACTOR VARIABLES PER MODEL

Tables 6 to 8 give, for every model, the analyses summarised in Section 5.2. Figure 3 averages over the model–task combinations whose P(Yes) AUROC exceeds 60 (Llama-3.2-3B, Phi-4-mini, Phi-4 and Qwen3-8B on PopQA; Phi-4-mini on GSM8K; the four Qwen3-VL models on DocVQA), so that the self-assessment the directions read carries signal; the tables report all models. In each, a

direction's projection is taken with the unanswerable pole positive, so a positive correlation means that the direction moves towards "unanswerable" as the variable increases. Table 6 uses $-\log$ of the Wikipedia page views of the subject entity of each PopQA question; Table 7 the number of non-empty lines of the GSM8K reference solution (2 to 11, median 3); Table 8 the Gaussian blur radius applied to the first 200 DocVQA images at 11 levels from 0 to 5 pixels, pooling all questions and levels.

Table 6: PopQA: Spearman correlation between direction projection and $-\log$ entity popularity, per model. Last row: correlation between the error label and the variable.

Direction	Llama-1B	Llama-3B	Phi-4-mini	Phi-4	Qwen3-0.6B	Qwen3-8B	Mean
Clarity	+0.460	+0.368	+0.500	+0.515	+0.070	+0.533	+0.408
Evidence	+0.342	+0.520	+0.521	+0.554	+0.010	+0.530	+0.413
Ambiguity	+0.511	+0.602	+0.491	+0.425	−0.082	+0.518	+0.411
Challenge	+0.007	+0.394	+0.426	+0.537	−0.025	+0.506	+0.307
Familiarity	**+0.601**	**+0.675**	**+0.606**	**+0.604**	−0.005	**+0.633**	**+0.519**
Certainty	+0.531	+0.602	+0.538	+0.550	**+0.071**	+0.558	+0.475
Feasibility	+0.232	+0.442	+0.554	+0.521	−0.006	+0.487	+0.372
Error label	+0.252	+0.320	+0.211	+0.340	+0.015	+0.252	+0.232

Table 7: GSM8K: Spearman correlation between direction projection and the number of reference-solution steps, per model.

Direction	Llama-1B	Llama-3B	Phi-4-mini	Phi-4	Qwen3-0.6B	Qwen3-8B	Mean
Clarity	−0.417	−0.322	+0.044	+0.036	−0.034	−0.208	−0.150
Evidence	−0.074	−0.152	+0.174	+0.109	+0.083	+0.089	+0.038
Ambiguity	−0.126	+0.141	+0.098	+0.094	+0.012	−0.109	+0.018
Challenge	**+0.082**	**+0.509**	**+0.477**	**+0.460**	**+0.296**	**+0.400**	**+0.371**
Familiarity	−0.124	−0.051	+0.172	+0.243	−0.091	−0.144	+0.001
Certainty	−0.201	+0.021	+0.125	+0.094	−0.087	+0.023	−0.004
Feasibility	−0.062	+0.155	+0.024	−0.121	+0.030	−0.285	−0.043
Error label	+0.178	+0.177	+0.113	+0.021	+0.288	+0.155	+0.155

Table 8: DocVQA blur: Spearman correlation between direction projection and blur radius over all 200 questions and 11 levels, per model (Q: Qwen3-VL, G: Gemma-3).

Direction	Q-2B	Q-4B	Q-8B	Q-32B	G-4B	G-12B	Mean
Clarity	+0.240	**+0.418**	+0.454	**+0.660**	**+0.437**	**+0.487**	**+0.449**
Evidence	+0.249	+0.227	+0.431	+0.589	+0.252	+0.280	+0.338
Ambiguity	**+0.314**	+0.403	+0.397	+0.623	+0.360	+0.443	+0.424
Challenge	+0.245	+0.111	+0.178	+0.439	+0.224	+0.316	+0.252
Familiarity	+0.199	+0.347	**+0.562**	+0.466	+0.304	+0.421	+0.383
Certainty	+0.227	+0.291	+0.487	+0.524	+0.328	+0.437	+0.382
Feasibility	+0.165	+0.343	+0.447	+0.453	+0.316	+0.327	+0.342

E Ablation

Table 9 changes one component of the VLM configuration at a time. The probing position matters most: the best of the other 19 candidate positions, the second token of the assistant suffix, is only three tokens before the last position and costs 4.31 points. Each direction contributes: dropping Feasibility costs 1.58 and dropping Challenge 2.69, so Challenge has the larger marginal contribution, and the combination beats either alone. Final-layer-only probing costs 1.00 (this also removes the Jacobian mappings), and replacing the Jacobian lens by the logit lens costs 0.40.

Table 9: Overall macro-average AUROC under five diagnostic changes to the final ADP configuration. Drops are relative to the full configuration.

Component	Variant	AUROC	Drop
ADP	Final configuration	73.75	–
Jacobian alignment	Logit lens	73.35	0.40
Multi-layer probing	Final layer only	72.75	1.00
Feasibility direction	Challenge only	72.16	1.58
Challenge direction	Feasibility only	71.06	2.69
Probing position	Best alternative position	69.44	4.31

F ROBUSTNESS

Irrelevant directions. Sixteen directions unrelated to answerability (Table 4) are added to Challenge and Feasibility with equal weight, enumerating all $\binom{16}{k}$ combinations for $k \leq 10$. The macro-average AUROC decreases monotonically from 73.75 to 70.56 at $k = 10$, about 0.3 points per added direction; about ten irrelevant directions are needed to erase the gain over ADP(Yes) (Figure 5).

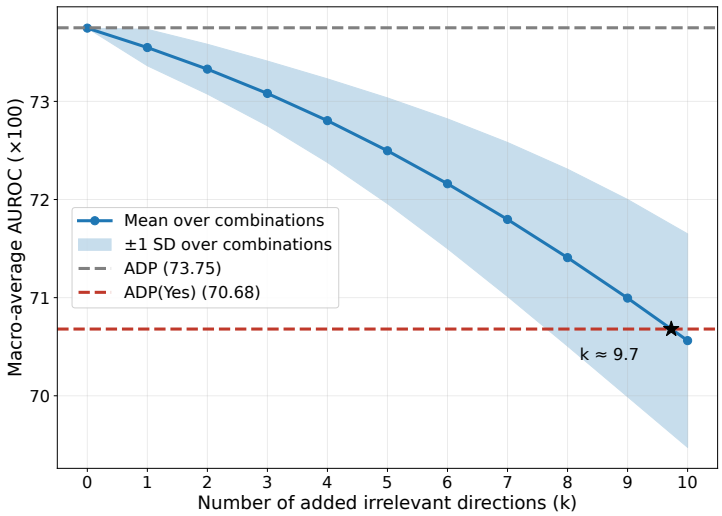


Figure 5: Overall macro-average AUROC as irrelevant directions are added to Challenge and Feasibility, showing mean ± SD over all combinations for each $k \leq 10$.

Synonym choice. Each pole of Challenge is extended from 4 to 6 words (*complex*, *daunting*; *straightforward*, *negligible*) and all $\binom{6}{4}^2 = 225$ four-word subsets are evaluated with Feasibility fixed. The macro average has mean 73.80, SD 0.10 and range 73.61 to 74.07; the reported set scores 73.75 and ranks 151st of 225, so it is not a favourable choice (Figure 6).

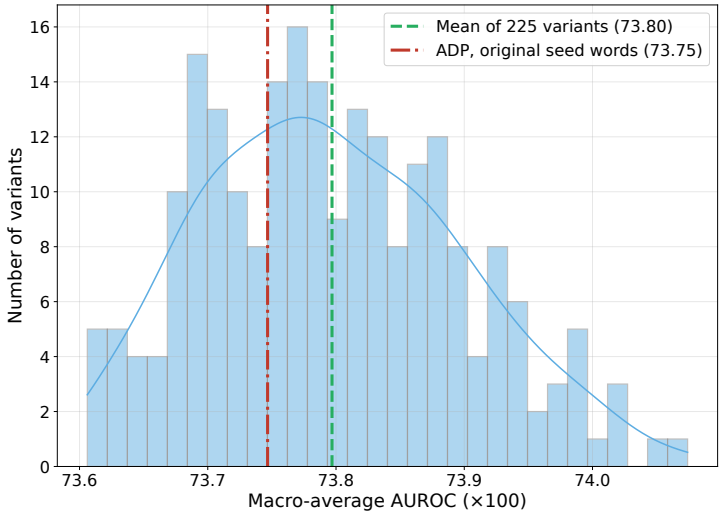


Figure 6: Distribution of overall macro-average AUROC over the 225 seed-word variants of the Challenge direction, with Feasibility fixed.

Random-word corruption and seed count. Replacing seed words by random words (from the five random directions) lowers the mean single-direction AUROC over the seven directions from 70.78 to 68.95 when half of the words are replaced, whereas adding the same number of random words lowers it only to 70.35 (Table 10, Figure 7). With approximately balanced subsets of the original $4 + 4$ words, the mean rises monotonically from 68.40 with two words to 70.78 with eight (Figure 8).

Table 10: Single-direction AUROC averaged over the seven answerability-related directions and the 18 model–task combinations when seed words are substituted with or supplemented by random words. Results are aligned by the fraction of random words in the seed set. AUROC is averaged over 10 random draws at each level, while Δ is reported as mean $\pm$ SE relative to the original seed set.

Perturbation	# words	Random frac.	AUROC	Δ
Substitution	4	50.0%	68.95	-1.83 ± 0.22
	3	37.5%	69.55	-1.23 ± 0.25
	2	25.0%	70.24	-0.54 ± 0.08
	1	12.5%	70.50	-0.28 ± 0.10
Original	0	0.0%	**70.78**	0.00
Addition	1	11.1%	70.62	-0.16 ± 0.07
	2	20.0%	70.59	-0.18 ± 0.04
	4	33.3%	70.49	-0.29 ± 0.05
	8	50.0%	70.35	-0.43 ± 0.05

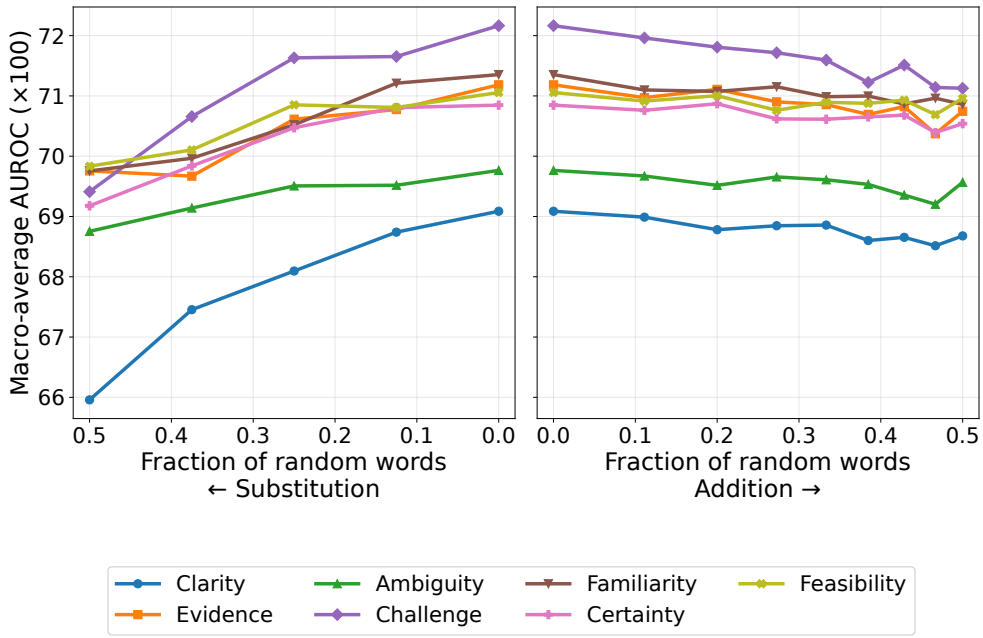


Figure 7: Single-direction AUROC for the seven answerability-related directions under seed-word substitution (left) and addition (right) with random words, as a function of the fraction of random words in the seed set.

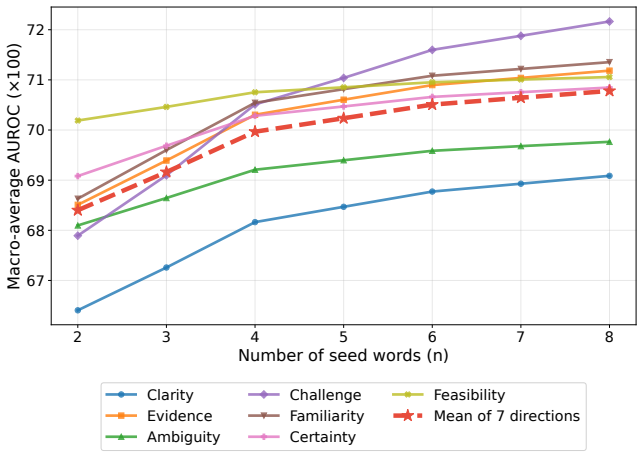


Figure 8: Single-direction AUROC as a function of the total number of seed words across the two poles ($n = 2$ to 8), using approximately balanced subsets of the original $4 + 4$ seed words and averaging over all valid subsets. All seven answerability-related directions and their mean are shown.

G Direction classes and the number of directions

Direction classes. The 25 directions fall into four classes (Table 11). Read alone, answerability-related directions average 70.78 AUROC over the 18 VLM combinations; control directions average 45.16 (four of six below 50), related 48.48 and random 48.21 (Table 12). Even after flipping every control direction that scores below 50 using the labels, a diagnostic that is not deployable, the answerability-related directions remain higher in every combination (Figure 9).

Table 11: The four direction classes and the properties that distinguish them.

Direction Class	Answerability relevance	Polarity	Pairing
Answerability-related	Yes	Yes	Yes
Control	No	Yes	Yes
Related	No	No	Yes
Random	No	No	No

Table 12: Mean single-direction AUROC by direction class, with the mean absolute deviation from 50. From Random upward, the adjacent class differences correspond to pairing, polarity and answerability relevance.

	Mean AUROC					
Direction Class	DocVQA	OK-VQA	MathVista	Qwen3-VL	Gemma-3	All
Answerability-related	**73.83**	**70.72**	**67.79**	**73.47**	**65.40**	**70.78**
Control	43.32	43.68	48.48	44.42	46.63	45.16
Related	47.36	48.76	49.34	48.03	49.39	48.48
Random	49.09	46.42	49.12	48.72	47.20	48.21
	Mean $\lvert$AUROC $-$ 50$\rvert$					
Direction Class	DocVQA	OK-VQA	MathVista	Qwen3-VL	Gemma-3	All
Answerability-related	**23.83**	**20.72**	**17.79**	**23.47**	**15.40**	**20.78**
Control	10.68	11.31	6.27	9.65	6.47	8.55
Related	6.20	3.83	2.33	6.13	5.51	3.48
Random	4.94	6.26	2.47	4.29	2.80	3.12

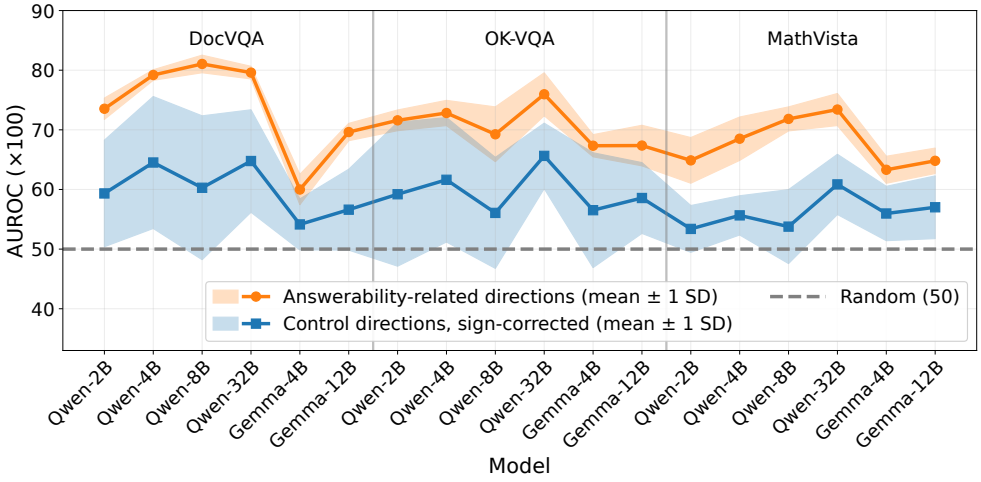


Figure 9: For each model–task combination, mean ± 1 SD of single-direction AUROC within the 7 answerability-related directions and within the 6 control directions. Control directions are sign-corrected post hoc using the labels and are therefore diagnostic only. SD is the population standard deviation across directions within each class.

Number of directions. All 127 non-empty subsets of the seven answerability-related directions are evaluated with equal weights on VLMs. The macro average rises from 70.78 with one direction to 72.40 with all seven, with positive gains in every combination; the second direction gives 53% to 77% of the total gain and the gain grows with model size within Qwen3-VL (Table 13, Figures 10 to 12). The two integrative directions reach 73.75, above the seven-direction average, which is the configuration used throughout.

Table 13: Macro-average AUROC across datasets at $k = 1$ and incremental gains as the number of answerability-related directions increases, averaged over all $\binom{7}{k}$ subsets at each k, by model. The *Total gain* row reports the gain from $k = 1$ to $k = 7$, and the k=2 share is the fraction of this total gain achieved from $k = 1$ to $k = 2$.

k	Qwen-2B	Qwen-4B	Qwen-8B	Qwen-32B	Gemma-4B	Gemma-12B
			Absolute AUROC			
1	70.00	73.51	74.04	76.32	63.54	67.27
			Gain over $k-1$			
2	**+0.968**	**+1.039**	**+1.219**	**+1.511**	**+0.539**	**+0.580**
3	+0.146	+0.263	+0.433	+0.576	+0.179	+0.226
4	+0.064	+0.120	+0.215	+0.311	+0.084	+0.110
5	+0.039	+0.069	+0.126	+0.196	+0.045	+0.062
6	+0.022	+0.044	+0.083	+0.137	+0.031	+0.039
7	+0.022	+0.034	+0.064	+0.098	+0.012	+0.023
Total gain	1.260	1.569	2.140	2.829	0.891	1.040
k=2 share	77%	66%	57%	53%	61%	56%

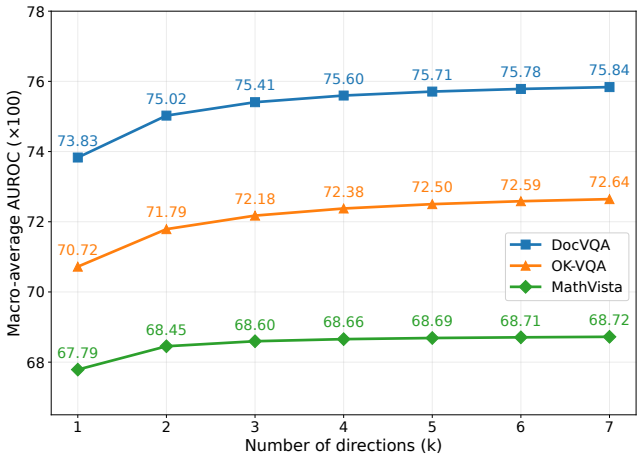


Figure 10: Macro-average AUROC of equal-weight combinations of k answerability-related directions by dataset, averaged over all subsets of size k.

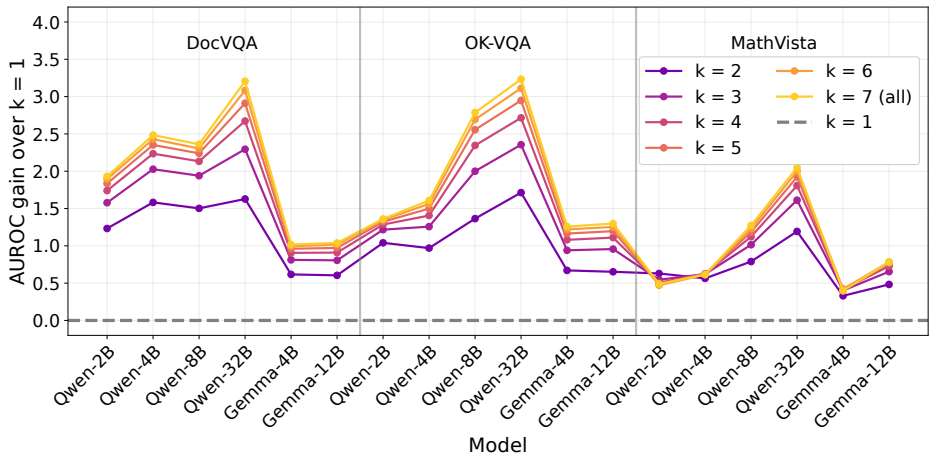


Figure 11: Gain in AUROC over $k = 1$ for each model–task combination as the number of answerability-related directions increases.

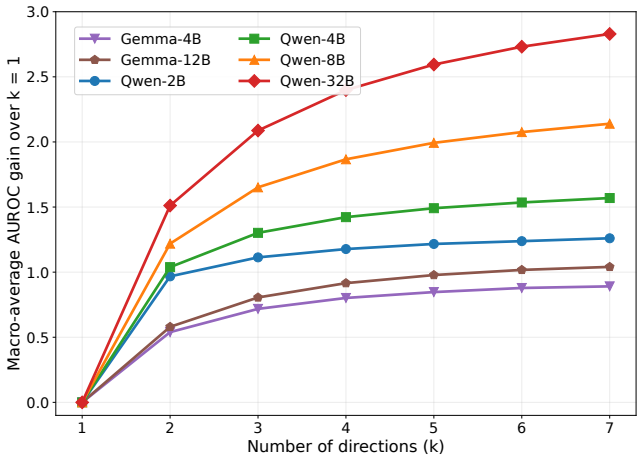


Figure 12: Gain in AUROC from $k = 1$ to $k = 7$, by model.

H LAYERS AND POSITIONS

Layers (VLMs). Probing a single layer at the last position, the AUROC averaged over the seven directions and 18 combinations rises with depth from 50.9 at 27% relative depth to a peak of 70.44, and averaging the projections over the whole probed span gives 70.78, above the best single layer, without choosing a layer from task data (Figure 13). The best position–layer cell of each direction and combination lies deep in the network (Figure 14).

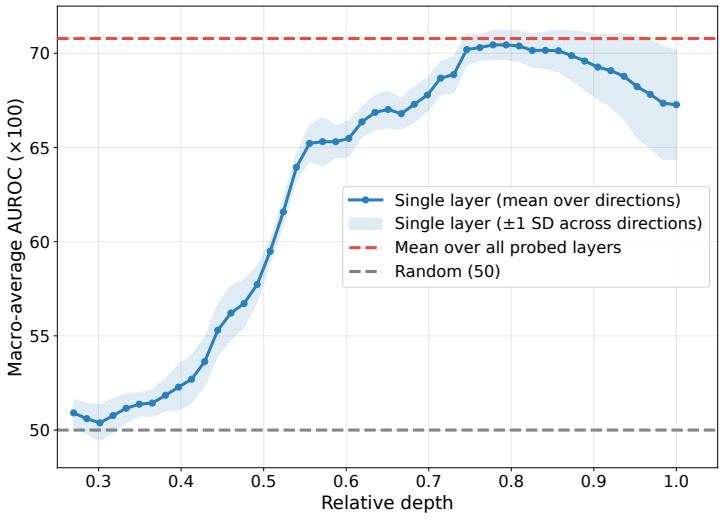


Figure 13: Single-layer probing AUROC at the last input position as a function of relative depth. Values are averaged over the 18 model–task combinations, and the line and band show the mean ± 1 SD across the 7 answerability-related directions. The dashed line shows the corresponding mean AUROC after averaging the probing scores over all probed layers.

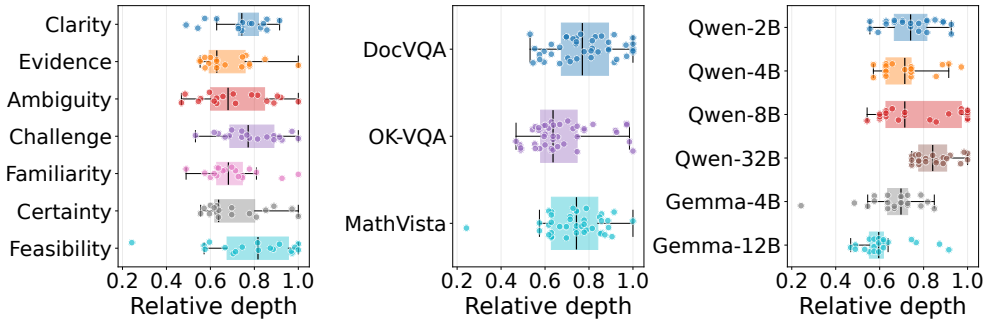


Figure 14: Relative depth of the best position–layer cell for each of the 7 answerability-related directions and 18 model–task combinations, giving 126 post-hoc maxima over the 20 positions and all probed layers, shown by model. These maxima are shown only to characterise where high-performing probing locations concentrate.

Positions (VLMs). Over the 20 candidate positions, the last input position is the best for all seven directions (mean 70.8); the next is the second suffix token (66.1). Position 11, the " No" token of the mapping line, reverses the sign of several directions, so intermediate positions are locally dominated by nearby words and do not read the same thing across inputs. Of the 126 post-hoc best position–layer cells, 78 lie at the last position and 105 within the suffix block (Figures 15 and 16).

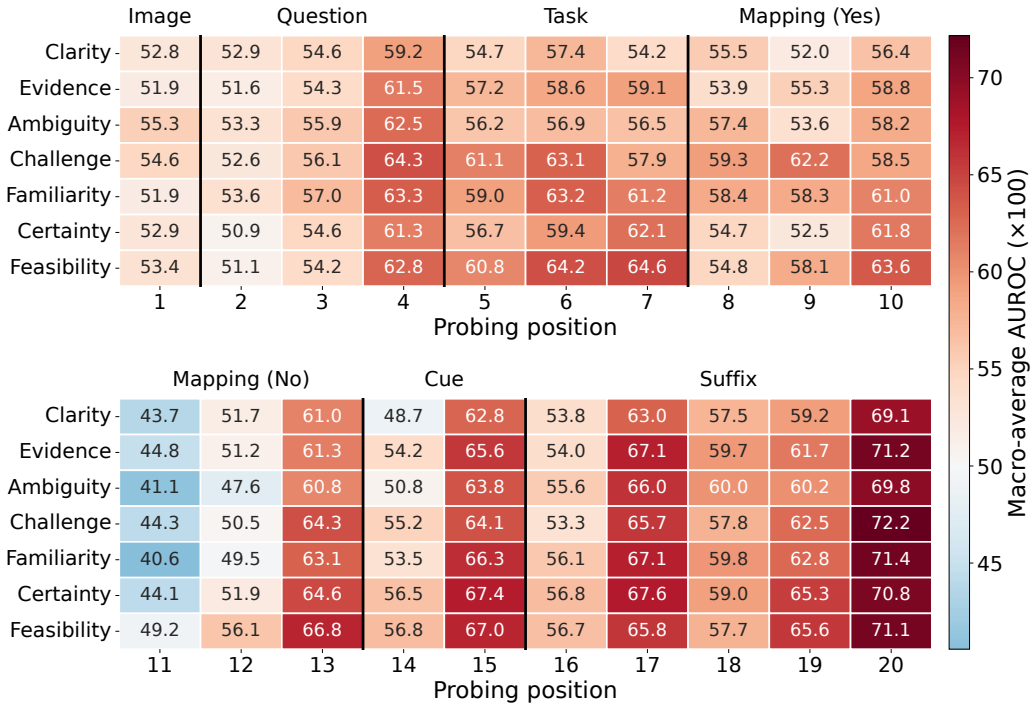


Figure 15: AUROC of layer-averaged probing at each of the 20 candidate positions for the 7 answerability-related directions, averaged over the 18 model–task combinations. Position 20 is the last input position, and position 11 is the " No" token of the mapping block.

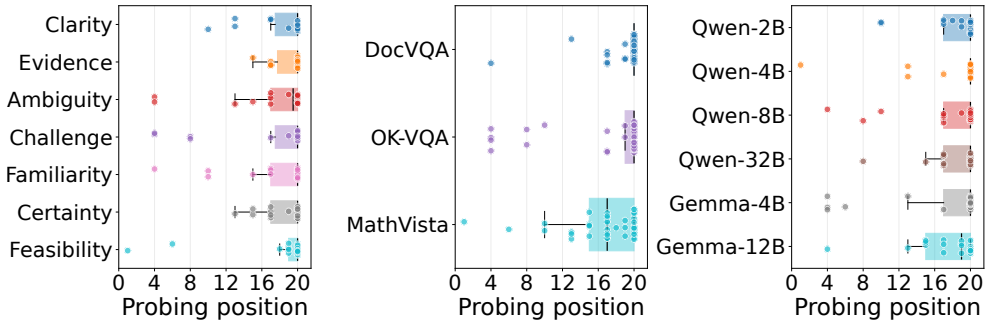


Figure 16: Position of the best position–layer cell for each of the 7 answerability-related directions and 18 model–task combinations, giving 126 post-hoc maxima over the 20 positions and all probed layers.

Positions (LLMs). For Qwen3 in non-thinking mode the last input positions are the assistant header and the empty think block that the chat template appends; reading each of the last six positions with the same ADP direction, the score is highest at the last token for both Qwen3-0.6B and Qwen3-8B, so the last input position is used for all models.

I Answer-format sensitivity of Yes/No probing

A score read from the *Yes*/*No* token pair can respond to the answer format as well as to answerability. Classifying questions by answer format (Table 14), MathVista has 18.9% binary questions against 5.1% for OK-VQA and 1.0% for DocVQA. On MathVista, P(Yes) falls from 61.65 on all questions to 52.22 on the binary subset and rises to 65.32 on the rest, for every model, with the binary drop growing from -5.5 for Qwen3-VL-2B to -20.4 for Qwen3-VL-32B (Table 15). Position–layer maps of P(Yes) show that on binary questions the informative cells move away from the final layer (Figures 17 to 19). ADP does not use the Yes/No direction and is not affected by this coupling.

Table 14: Answer-format composition of the three datasets, classified by Qwen3-VL-32B from the question and reference answer. A denotes binary questions with exactly two possible answers, B denotes multiple-choice questions with more than two options, and C denotes free-form questions.

Dataset	Binary (A)	Multi-choice (B)	Free-form (C)	Total
DocVQA	56 (1.0%)	27 (0.5%)	5266 (98.4%)	5349
OK-VQA	259 (5.1%)	27 (0.5%)	4760 (94.3%)	5046
MathVista	**189 (18.9%)**	21 (2.1%)	790 (79.0%)	1000

Table 15: AUROC of P(Yes) on the full MathVista test set and its binary and non-binary subsets. Δ is relative to the full set.

Model	All	Binary		Non-binary	
		AUROC	Δ	AUROC	Δ
Qwen-2B	60.05	54.53	-5.52	64.92	$+4.88$
Qwen-4B	62.13	55.11	-7.02	66.34	$+4.20$
Qwen-8B	66.90	48.70	-18.20	71.77	$+4.87$
Qwen-32B	64.46	44.05	-20.41	68.14	$+3.68$
Gemma-4B	57.98	54.59	-3.39	60.29	$+2.32$
Gemma-12B	58.35	56.34	-2.01	60.43	$+2.07$
Mean	61.65	52.22	-9.43	65.32	$+3.67$

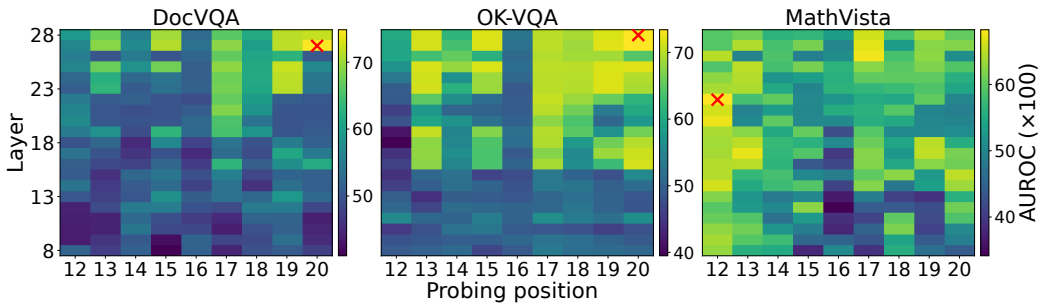


Figure 17: P(Yes) AUROC at each of the last nine candidate positions (12–20) and each evaluated layer for Qwen3-VL-2B on the three datasets. The cross marks the best cell within the displayed range, and the top-right cell is the fixed P(Yes) point.

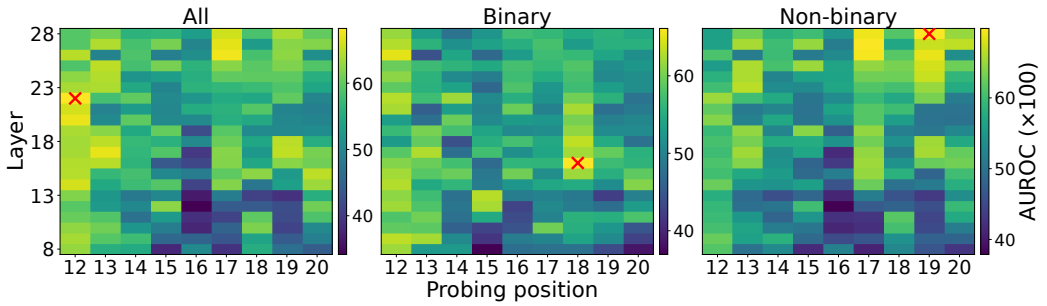


Figure 18: P(Yes) AUROC across positions and layers on the binary and non-binary MathVista subsets for Qwen3-VL-2B. The cross marks the best cell within the displayed range.

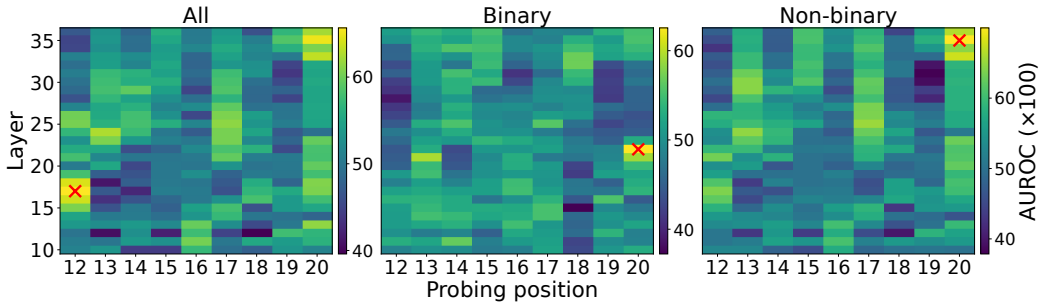


Figure 19: P(Yes) AUROC across positions and layers on the binary and non-binary MathVista subsets for Qwen3-VL-4B. The cross marks the best cell within the displayed range.