A Linear Answerability Signal That the Model Does Not Act On

Probes, verbal confidence and an abstain option in Phi-3.5-mini

Mackenzie O'Brien, Mac Labs AI

November 2026

AI x Epistemics Research Sprint · Track 1: model epistemics and decision-making evals

The question: can the model tell, and does it act on it?

Why it matters

- A model that answers every question in the same confident tone leaves its reader unable to tell which answers to trust.
- The problem is sharpest when the evidence needed for an answer is missing.
- This is an evaluation question: can a model's expressed confidence be checked against an internal signal and against ground truth about answerability?

What we ask

- When Phi-3.5-mini-instruct reads a passage and a question the passage cannot answer, does its hidden state carry a *linearly readable* trace of that fact?
- Does anything the model *says or does* (verbal confidence, token statistics, sampling agreement, an abstain option) reflect the trace?

Scope

One small model, one run, one prompt. A supervised probe shows a signal exists. It does not show the model uses it.

Setup: model, data and splits

Model

- `microsoft/Phi-3.5-mini-instruct`: 3.8B parameters, 32 layers, hidden size 3072
- 8-bit quantization on one 8 GB GPU, batch size one (the bf16 weights do not fit)
- Greedy decoding, at most 32 new tokens; one run, seed 13

Data

- **SQuAD 2.0** (primary): 600 answerable and 600 unanswerable dev questions, split by passage 60/20/20 (720/240/240)
- **SelfAware**: 1,000 per class, split by question 1,200/400/400
- Pooled: **3,200 items** (1,920/640/640)
- SQuAD test: n=240, 120 per class

Three model calls per item

1. **Plain prompt**: answer, token log-probabilities and entropy, and the hidden states we probe
2. **Confidence turn**: "How confident are you that your answer is correct?" on 0–100
3. **Explicit abstain option**: plain prompt plus "If the question cannot be answered, reply exactly UNANSWERABLE."

A hash-selected 25% of items (807 of 3,200; 51 of the 240 SQuAD test items) also get five samples at temperature 0.7 for a consistency signal.

Scoring: SQuAD answer correct on exact match or token-F1 ≥ 0.5; abstention found by a rule list; no judge model.

Method: probes, baselines and what was chosen where

Probe

- Hidden state at the *last prompt token*, the position that generates the first answer token, for all 33 residual positions (embedding plus 32 layers)
- Standardized logistic regression ($C = 1$), trained on the training split only; a mass-mean probe as a robustness check
- Layer picked by **validation** AUROC

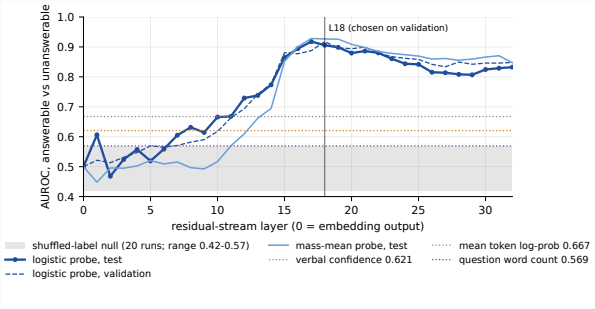
Compared against

- Verbalized confidence, mean token log-probability, negative token entropy, sampling consistency, the abstain option
- Controls: 20 shuffled-label runs, a question-length baseline, post hoc lexical checks

Choice	Made on	Reported on
Probe layer (L18 for SQuAD), by AUROC	validation	test
Probe and confidence thresholds, by Youden's J	validation	test
Regularization $C = 1$ and confidence cut 0.8	fixed in config	test
Transfer layer, "reported"	source validation	target test
Transfer layer, "max over layers"	target test (optimistic)	target test

Intervals are 95% percentile bootstrap intervals over test items (300 resamples), ignoring clustering by passage. One seed, one split.

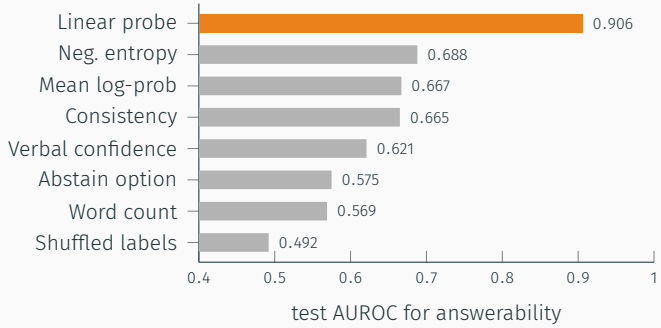
Result: a linear signal at middle layers



SQuAD 2.0, n=240 per split. Shaded band: range of 20 shuffled-label runs. Vertical line: layer chosen on validation.

- Test AUROC **0.906** (95% CI 0.867–0.948) at **layer 18**, the best validation layer (validation 0.918).
- Through layer 11 the test AUROC stays at or below 0.668; it reaches 0.865 at layer 15, peaks at 0.918 at layer 17 and falls to 0.832 at layer 32.
- Layer 0 scores exactly 0.500, as it must when every item ends in the same token.
- Mass-mean probe picks the same layer: 0.927 on test (validation 0.912).
- At the validation threshold: balanced accuracy 0.821.
- The best test layer (17) is only 0.012 above the reported value, so selection on test data does not inflate the number.

Baselines: the model's own signals do much worse



SQuAD 2.0 test, n=240. Consistency covers only 51 test items. The abstain option is one hard decision, so its AUROC equals its balanced accuracy.

- Verbalized confidence: **0.621** [0.558, 0.681].
- Token-level and sampling signals: 0.665–0.688.
- Question length: 0.569 (word count), 0.558 (characters).
- Shuffled-label probes: mean **0.492** (SD 0.039, range 0.418–0.568, none above 0.8), so selecting a layer on 240 validation items does not by itself produce a high score.
- Post hoc lexical checks: TF-IDF of the question 0.598, question–passage word overlap 0.600, both 0.635.
- These baselines do not exclude other surface cues, such as how crowdworkers wrote the unanswerable questions.

Behavior: the model rarely abstains, with or without an option

Plain prompt (no way to decline)

- Answerable test questions answered correctly: 90.0%
- Abstains on **6 of 120** unanswerable items (5.0%), and on none of the answerable ones
- Action score 0.525 against recognition score 0.821: a recognition–action gap of **0.296** (post hoc 95% CI 0.240–0.344)

With the explicit UNANSWERABLE option

- Abstains on **18 of 120** unanswerable items (15%), and on none of the answerable ones
- Precision 1.0 (18 of 18), recall 0.150 (post hoc CI 0.091–0.208), F1 0.261
- Balanced accuracy 0.575; the gap becomes 0.246 (CI 0.189–0.292)

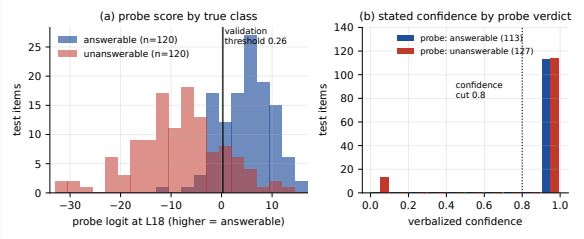
How to read this

The option is conservative rather than inaccurate: it never refuses an answerable question, but it leaves 102 of 120 unanswerable questions answered.

Caveats

We tried one wording of the option. The plain prompt offers no way to decline, so a low action score is expected of any model that follows the instruction.

The behavioral gap: the probe flags it, the model commits anyway



(a) Probe logit at layer 18 by true class; the line is the validation threshold. (b) Verbalized confidence by probe verdict.

- The probe flags **127 of 240** test items as unanswerable (102 truly so, 25 answerable).
- Of these, 112 were answered without abstaining at confidence ≥ 0.8: **46.7%** of all test items (post hoc CI 41.0–52.9%).
- 92 of the 112 were answered wrongly: an error rate of **0.821** (CI 0.747–0.888), against 0.525 over all 240 items and 0.266 outside the gap.
- Among the 120 unanswerable items, 87 (72.5%) fall in the gap.

Two limits: the confidence condition barely filters (225 of 240 items are answered at confidence ≥ 0.8), and the gap only shows that a probe trained on 720 labeled items recovers a distinction the model does not act on under these prompts. It does not show the model "withholds" anything.

Flagging unanswerable items: probe, confidence and abstain option

Rule	Recall	Precision	F1	Coverage	False abstention
Probe, L18	**0.850**	0.803	**0.826**	0.471	0.185
Verbal confidence < 0.95	0.200	0.828	0.322	0.879	0.037
Explicit abstain option	0.150	**1.000**	0.261	0.925	0.000

SQuAD test, n=240 (120 unanswerable). Probe and confidence thresholds were fitted on validation labels; the abstain option used no labels. False abstention: share of correctly answered answerable items the rule would refuse.

- Flagging at the validation threshold keeps 47.1% coverage and raises accuracy among answered items from 0.450 to 0.779.
- The price: it refuses 18.5% of correct answers.
- A cheap gate needs white-box access and labeled training data.
- The comparison favors the probe and confidence rules, because they used validation labels and the option did not.

A source confound: SelfAware looks perfect and is not evidence

- Probe on **SelfAware**: AUROC **0.993** [0.985, 0.998] at layer 15 (test n=400).
- But a classifier on the **source label alone** reaches **1.000**: all 1,000 unanswerable items have source `SelfAware`, and none of the 1,000 answerable items does.
- A TF-IDF classifier on the question text reaches 0.923: question style alone separates the classes.
- Pooled probe: 0.959 [0.944, 0.972] at layer 18 (n=640), where a source-only feature gives 0.930.
- **Transfer is weak**: SelfAware→SQuAD 0.580 (layer 15); SQuAD→SelfAware 0.671 (layer 18).

SelfAware source	Answerable	Unanswerable
`SelfAware`	0	1000
`hotpot_dev`	6	0
`hotpot_train`	64	0
`squadqa_dev`	71	0
`squadqa_train`	551	0
`triviaqa_dev`	4	0
`triviaqa_train`	304	0

We report SQuAD 2.0 as the primary result and SelfAware as a documented confound. Probe results on benchmarks built from mixed sources need a source audit first.

What we claim, and what we do not

The results establish

- In one 3.8B model run in 8-bit, with one prompt format, a linear readout of the last prompt token at layers 15 to 19 predicts SQuAD 2.0 answerability on held-out passages (AUROC 0.906, CI 0.867–0.948, n=240).
- That is far above a shuffled-label null (mean 0.492), a question-length baseline (0.569) and the signals the model expresses (0.621–0.688).

They do not establish

- That the model represents “not knowing”.
- That it uses the signal, or that the signal is causal (we made no intervention).
- That the signal transfers across datasets or generalizes to other models.
- The SelfAware and pooled numbers are explicitly confounds; the numbers come from one seed and one split.

Practical reading

A probe is a cheap gate for white-box users (flags 85% of unanswerable items at 80% precision) but not a certificate that an answer is supported.

Limitations

- **Statistical power**: the SQuAD test split has 240 items (120 per class); only 12 answerable items were answered wrongly; the consistency signal covers 51 items.
- All comparisons reuse the same test items, with no multiple-comparison correction and no paired tests; bootstrap intervals ignore passage clustering.
- **Measurement**: no judge audited the rule-based abstention detector; SelfAware answers are scored by string containment, which understates accuracy.
- The probe sees only the plain prompt, while the abstain-option pass uses a different prompt, so the two are not evaluated on identical inputs.
- Prompts were truncated at 1,024 tokens (not logged how many); 8-bit quantization may shift hidden states.
- SQuAD's unanswerable questions were written adversarially, so surface artifacts beyond those tested may exist.
- The abstain-option analysis code was added after the run, and its numbers were recomputed from the run's saved responses.

What's next

Most informative next steps

- An **interventional test**: add or ablate the probe direction and measure abstention, in the manner of Lavi et al.
- A prompted answerability question and a P(True) baseline.
- Several abstain-option wordings.
- Multiple seeds and splits, and a second model.
- An unconfounded inherent-unanswerability dataset in which source does not alias the label.

Links and references

- **Code (GitHub):** `https://github.com/mso-docs/Apart-Epistemics`
- **Dataset (Hugging Face):** `https://huggingface.co/datasets/mackenzietechdocs/apart-epistemics-phi35-abstention`
- **Landing page:** `https://mso-docs.github.io/Apart-Epistemics/`
- **Data:** SQuAD 2.0 and SelfAware are public (CC BY-SA 4.0).

Key references (arXiv IDs as listed in the write-up): Phi-3 technical report (2404.14219); SQuAD 2.0, Rajpurkar et al. (1806.03822); SelfAware, Yin et al. (2305.18153); geometry of truth probes, Marks and Tegmark (2310.06824); control tasks for probes, Hewitt and Liang (1909.03368); curious case of hallucinatory (un)answerability, Slobodkin et al. (2310.11877); linear directions for (un)answerability, Lavi et al. (2509.22449).

Author: Mackenzie O'Brien, Mac Labs AI (sole author). Research conducted at the AI x Epistemics Research Sprint, November 2026.