

Uncertainty and Calibration in Large Models

Reliable Reasoning and Agent Safety

Wenbo Hu

Hefei University of Technology

CNCC 2026 Tutorial Chengdu 21 October 2026

Outline and Module Scope

- 01** Foundations of Uncertainty and Calibration
- 02** Uncertainty Measurement in Large Models
- 03** Uncertainty in Chain-of-Thought Reasoning
- 04** Over-Refusal and Safety Boundaries
- 05** Evidence Alignment for Agent Safety

This module: confidence diagnosis, internal evidence and safety boundaries.

Next: uncertainty-guided RL, confidence learning and external verification.

Uncertainty in Large Model Predictions

Illustrative question

“What is the capital of Australia?”

Answer: “Sydney.” Reported confidence: 95%

The target event

c = estimated probability that this answer is correct

Accuracy measures outcomes across examples. Confidence also needs validation.

Teaching example. Calibration background: Guo et al., ICML 2017.

Aleatoric and Epistemic Uncertainty

Source	Illustrative situation	Useful response
Aleatoric uncertainty	Ambiguous question or incomplete context	Clarification or a distribution over plausible answers
Epistemic uncertainty	Missing knowledge or a poorly learned relationship	Additional evidence or model improvement

The distinction depends on the information and modeling assumptions.

Sampling diversity can reflect paraphrases as well as uncertainty.

Background: Kendall and Gal, NeurIPS 2017. Examples are illustrative.

Calibration and Error Ranking Are Distinct

$$\Pr(Z = 1 \mid c = q) = q \iff \mathbb{E}[Z \mid c] = c$$

Property	Question
Accuracy	How often is the answer correct?
Calibration	Does the confidence match the observed frequency?
Discrimination	Does the score rank correct answers above errors?

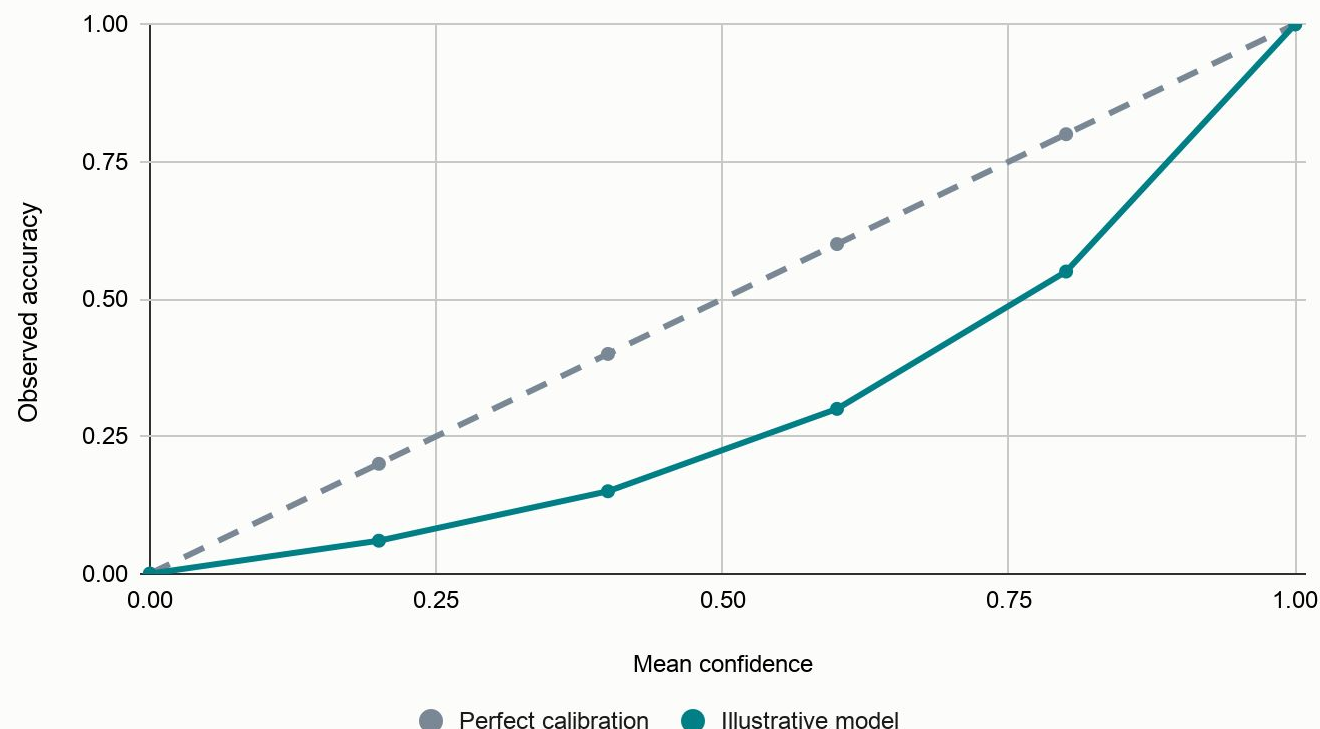
INSIGHT 1 Calibration does not guarantee useful error ranking.

A constant base-rate score can be calibrated yet cannot identify errors.

Z denotes answer correctness. Background: Guo et al., ICML 2017.

Reliability Diagrams and Calibration Metrics

Illustrative reliability



Expected calibration error

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} \cdot |\text{acc}(B_m) - \text{conf}(B_m)|$$

Binary Brier score

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (c_i - Z_i)^2$$

Equal-bin ECE = 0.20
B denotes a confidence bin.

Synthetic values for teaching. ECE depends on binning and finite-sample variation.

Ranking and Cost Guide Selective Answering

Answer when $c \geq \tau$

Threshold τ	Coverage	Error among answers
0.50	100%	20%
0.70	80%	15%
0.90	40%	5%

Expected loss of answering: $(1 - c) \times \text{error cost}$

Coverage = $\Pr(c \geq \tau)$ Selective risk = $\mathbb{E}[1 - Z \mid c \geq \tau]$

Illustrative thresholds. Coverage is the answered fraction. Abstention and verification also carry costs.

Measuring Uncertainty in Large Models

Signal	Operational score	Access and main caveat
Token likelihood	$\exp\left(\frac{1}{T} \sum_{t=1}^T \log p(y_t \mid y_{<t}, x)\right)$	Logits required Wording and length effects
Verbal confidence	Model-reported probability	Text API External calibration needed
Semantic entropy	$H = -\sum_k q_k \log q_k$ q: meaning-cluster frequency	Repeated answers + clustering Consistent errors can have low H
Internal probe	Score from selected hidden states	White-box access + labels Test under distribution shift

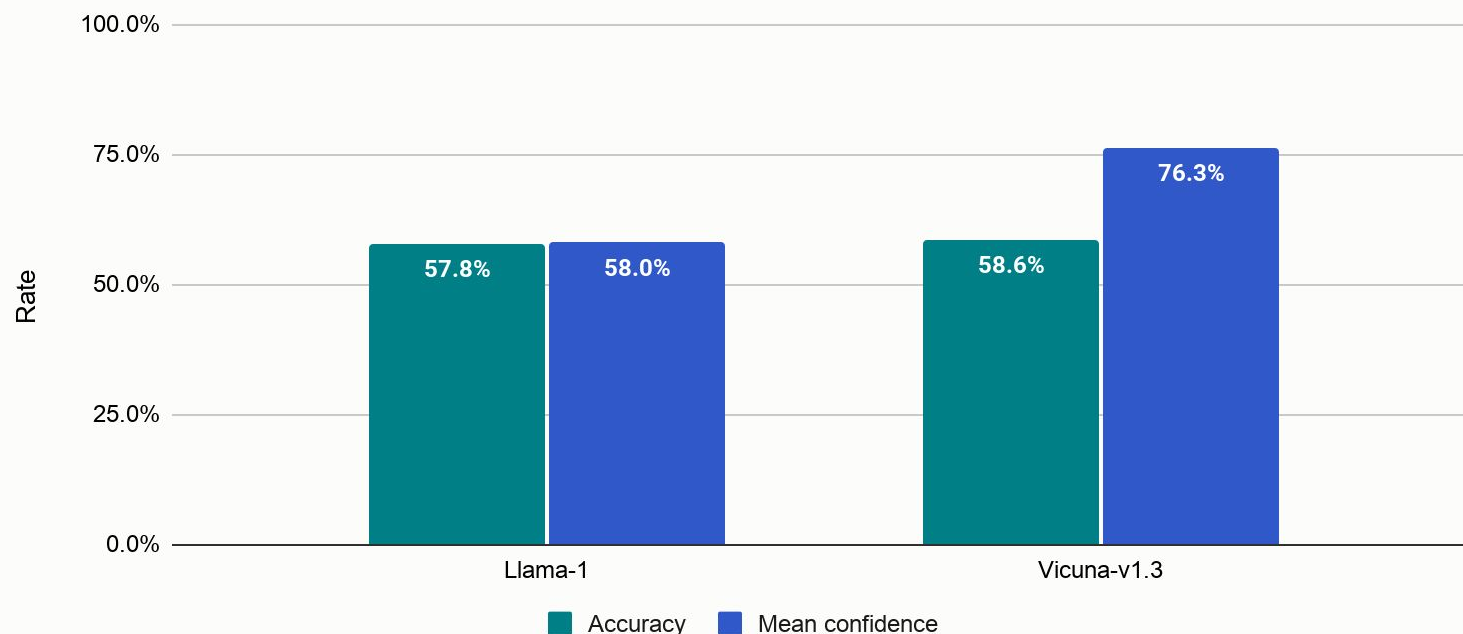
3 paraphrases: $H = 0$ **3 equally frequent meanings:** $H = \ln 3$

Background: He et al., 2023; semantic entropy, Farquhar et al., Nature 2024; DHCR, AAAI 2026.

Confidence Can Shift More Than Accuracy

MMLU, 5-shot, 33B models from the 2023 study

MMLU, 5-shot, 33B



Accuracy

+0.87 pp

Mean confidence

+18.26 pp

He et al., Investigating Uncertainty Calibration of Aligned Language Models, 2023, Figure 1.

Format Changes What Confidence Measures

$$p(y \mid x) = p(y \mid x, F) \times p(F \mid x)$$

Event	Illustrative probability	Interpretation
Use choice-letter format F	0.60	Format preference
Answer A, given F	0.80	Conditional answer preference
Answer A in format F	0.48	Joint response probability

INSIGHT 2 Confidence reflects answer belief and response behavior.

Define the scoring event before comparing prompts or models.

He et al., 2023, Section 4.1. Unique-format assumption. Numerical example is illustrative.

Calibration Depends on the Alignment Target

Llama-1 7B: synthetic MCQ alignment, MMLU validation

Scheme	Optimization target	MMLU calibration outcome
Pre-trained + ICL	Reference, without alignment	Reference confidence and ECE
Format only	SFT/DPO on response format	Confidence and ECE stay near reference
Choice or mixed	SFT/DPO including answer choice	Confidence and ECE increase

INSIGHT 2 Calibration changes depend on what alignment optimizes.

Format-only and answer-choice training have different effects in this experiment.

He et al., 2023, Section 4.3, Figure 7. Qualitative summary of controlled SFT/DPO schemes on a synthetic task.

Calibration Transfer Varies by Backbone

MMLU: text-only evaluation after multimodal training

Model	Accuracy	Mean confidence	ECE
Vicuna-7B	37.5%	58.9%	0.213
LLaVA-7B	42.1%	64.0%	0.224
Vicuna-13B	41.4%	66.7%	0.253
LLaVA-13B	43.1%	73.9%	0.308
Qwen-7B	42.4%	55.0%	0.134
Qwen-VL-Chat	50.0%	59.9%	0.099

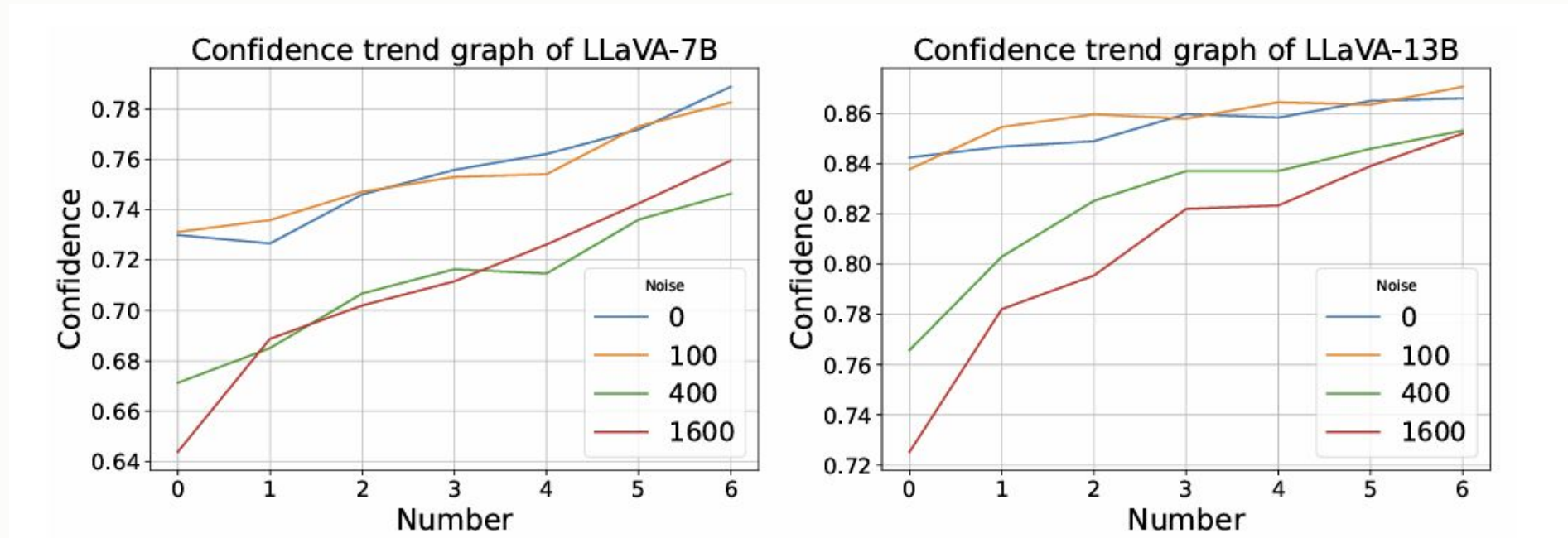
LLaVA pairs: ECE increases. Qwen pair: ECE decreases.

Revalidate calibration after changes to the backbone or training regime.

Chen et al., COLING 2025, Table 2. Selected matched pairs. These are language-task results without image inputs.

Text Can Raise Confidence on Noisy Images

Confidence rises as more textual description is supplied.



Horizontal axis: number of added description sentences.
Confidence changes alone do not establish improved accuracy.

Chen et al., COLING 2025, Figure 3 and Appendix C. Curves correspond to different image-noise levels.

IDK Prompts Change Uncertainty Admissions

Rate of acknowledging uncertainty on the designated OOD questions

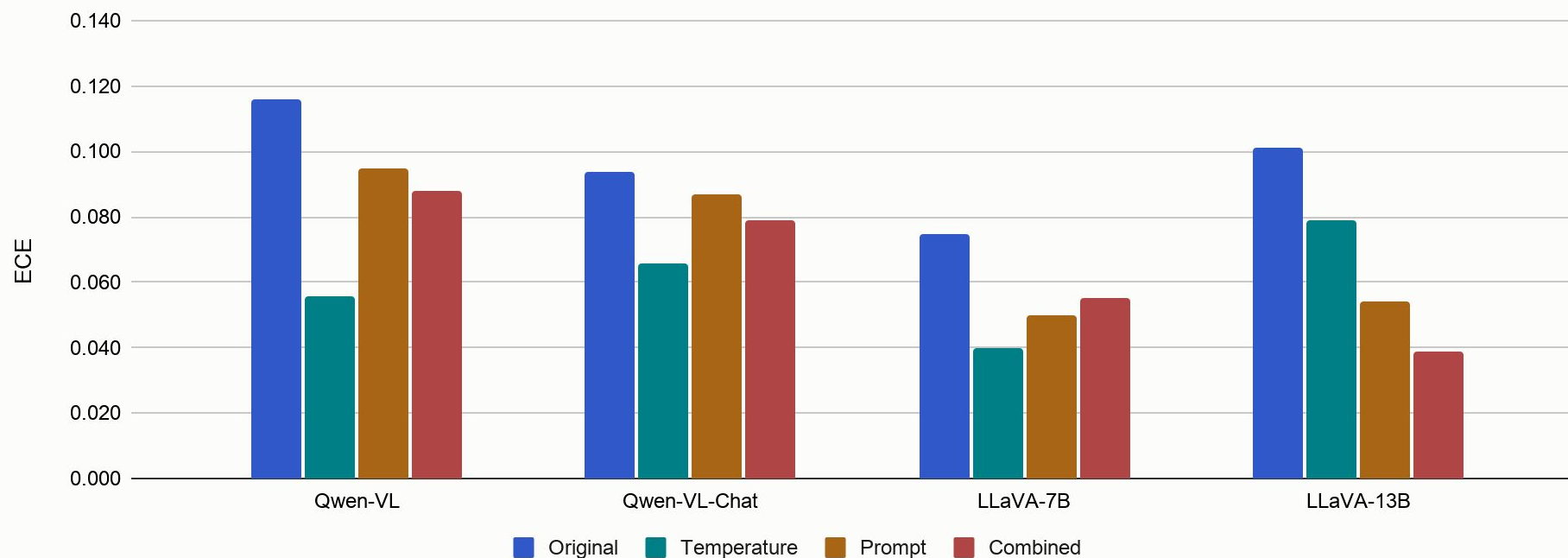
Model	Default	With IDK prompt
GPT-4o	6.1%	69.7%
Claude-3-Haiku	13.3%	67.7%

INSIGHT 2 REVISITED Admissions of uncertainty are prompt-sensitive.
More IDK responses alone do not establish better calibration.

Chen et al., COLING 2025, Table 4. 1,000 sampled July24-NewsVQA questions. Post-release date is a knowledge-gap proxy.

Calibration Effects Depend on the Model

MMBench calibration



ECE ↓

MMBench

Discrete text
prompt
optimization

The best method varies by model.

Chen et al., COLING 2025, Figure 4 and Algorithm 1. ECE is dimensionless.

Uncertainty in Chain-of-Thought Reasoning

Example: start with 3 apples, receive 2, then give away 1

Step	Valid continuation	Faulty continuation
Receive 2	$3 + 2 = 5$	$3 + 2 = 6$
Give away 1	$5 - 1 = 4$	$6 - 1 = 5$

Next-step validity and final-answer correctness are different events.

Illustrative arithmetic example. Research context: DHCR, AAAI 2026.

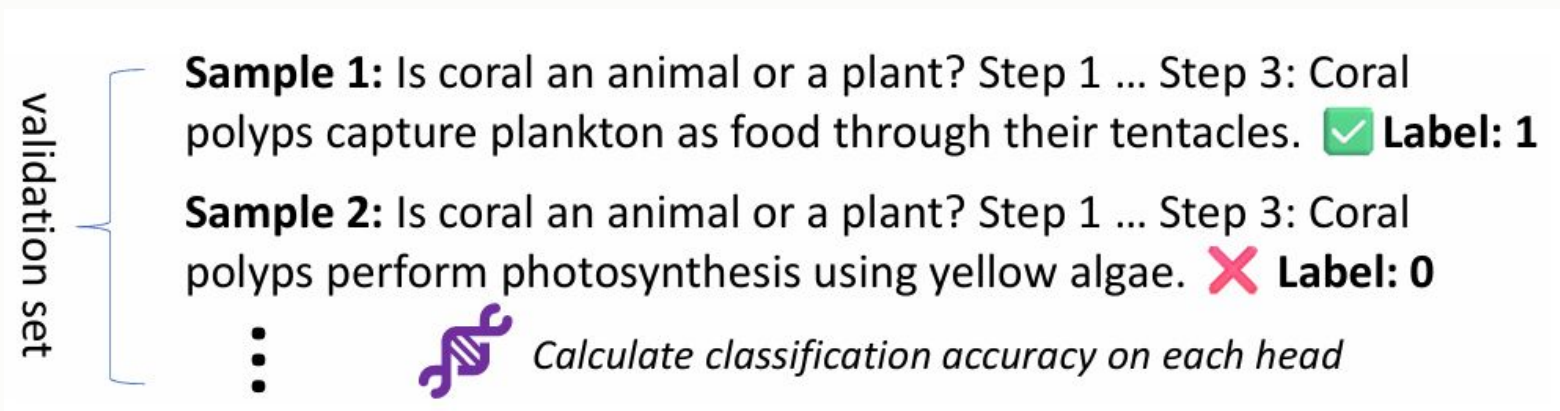
Selecting Among Reasoning Paths

Approach	Selection signal	Main limitation
Self-consistency	Final-answer vote	Agreement may share errors
Self-evaluation	Model judges a candidate	Judge reliability needs checking
Process reward model	Learned external step evaluator	Supervision and distribution shift
DHCR	Selected internal activations	Hidden-state access and labels

Sampling budget and inference cost belong in the comparison.

Wang et al., ICLR 2023; Chen et al., DHCR, AAAI 2026.

Correctness Is Decodable from Hidden States



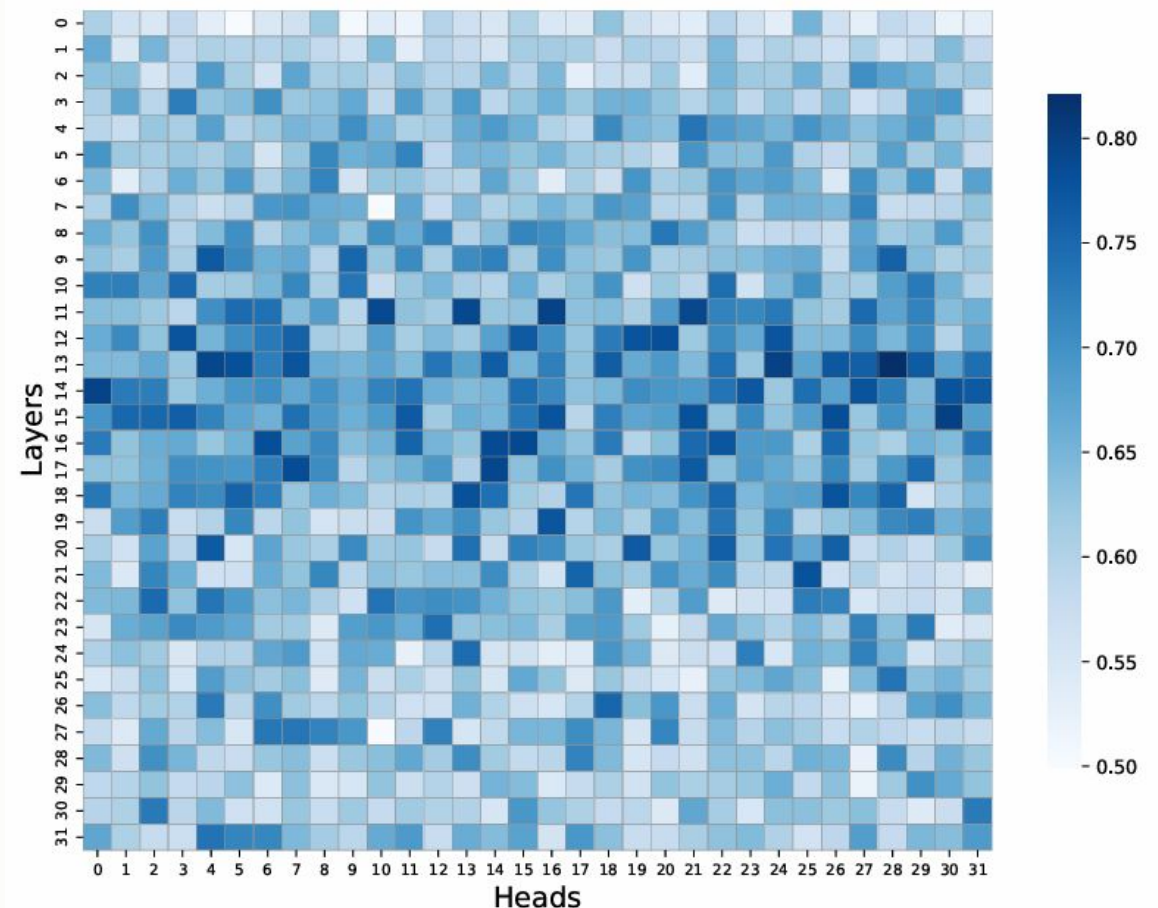
INSIGHT 3 Hidden correctness evidence can disagree with likelihood.

Use held-out correctness labels to test what the hidden state predicts.

Decision implication: test whether this signal improves candidate selection.
Predictive recoverability alone does not establish a causal mechanism.

DHCR, AAAI 2026, constructed CoT dataset and probing analysis.

Correctness Evidence Varies by Head and Layer



One cell, one attention head

Color shows correctness-classification accuracy.

Predictive heads

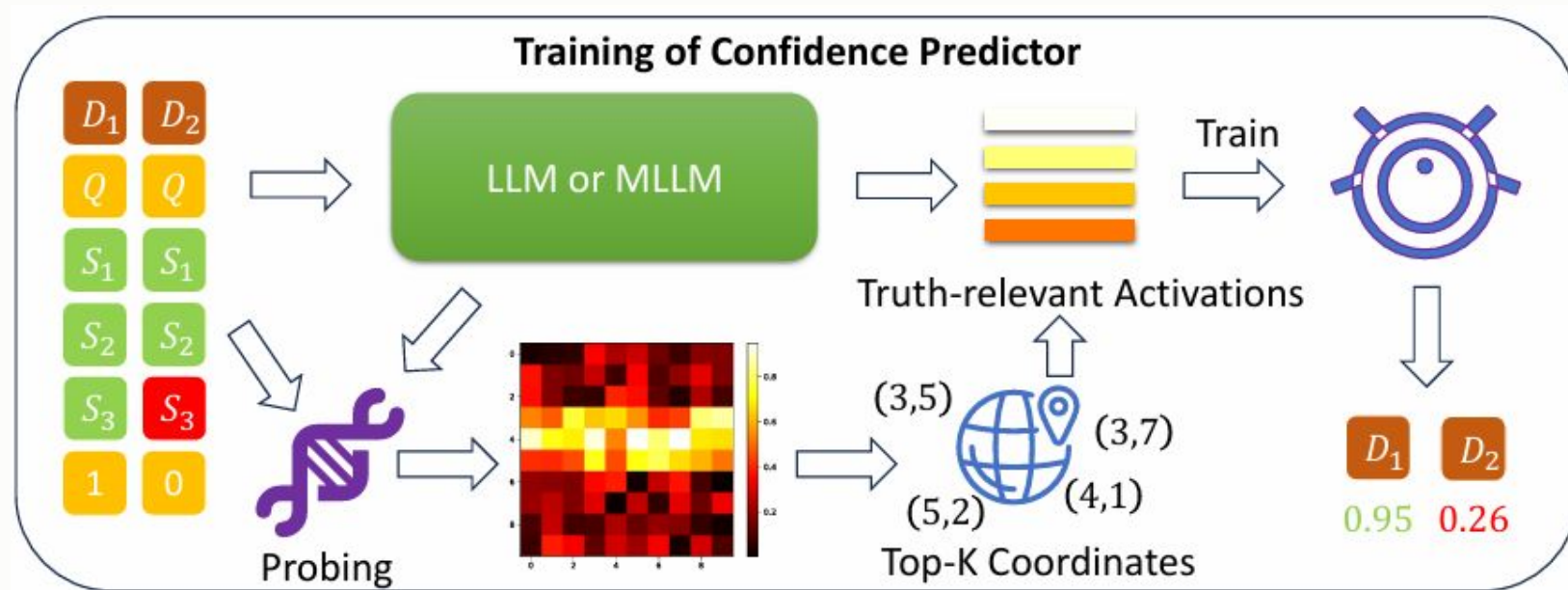
Retain informative heads selected on validation data.

Held-out evaluation

Feature selection and final testing use separate data.

DHCR, AAAI 2026, probing section. Truth-sensitive means predictive of the evaluated labels.

DHCR: Calibrating Step-Level Confidence



Selected activations

Concatenate features and fit a sigmoid predictor.

$$\beta = \sigma(\mathbf{w}^T \mathbf{v} + b)$$

Training uses cross-validated bin-accuracy soft targets.

DHCR, AAAI 2026, Equations 3–4. The ECE-inspired training objective differs from reporting ECE.

Calibration and Ranking Can Disagree

LLaMA2-13B-Chat; separate answer-correctness benchmarks

Dataset	DHCR ECE	ACTCAB ECE	DHCR Brier	ACTCAB Brier	DHCR AUROC	Seq. LL AUROC
WikiQA	0.037	0.058	0.102	0.149	0.934	0.678
TruthfulQA	0.019	0.038	0.122	0.140	0.907	0.480
SciQ	0.018	0.052	0.095	0.132	0.937	0.949
CommonQA	0.024	0.029	0.133	0.137	0.891	0.873

INSIGHT 1 REVISITED Calibration and ranking can favor different scores.

SciQ: DHCR ECE 0.018; sequence likelihood has higher AUROC (0.949 vs 0.937).

DHCR, AAAI 2026, Table 1. ECE and Brier lower is better; AUROC higher is better.

DHCR: Likelihood Can Favor an Incorrect Step

Candidate continuation	Generation probability	Predicted correctness	Combined score
Coral polyps capture plankton as food.	0.69	0.85	0.770
Coral polyps perform photosynthesis.	0.89	0.46	0.675

$$\text{Score}(C) = \lambda \beta(C) + (1 - \lambda) \bar{P}(C)$$

$\bar{P}$ is length-normalized generation probability. Here $\lambda = 0.5$

DHCR, Figure 1 and Equation 5. Weighted scores are a teaching calculation from the illustrated candidate values.

DHCR Gains Vary by Baseline and Model

Average accuracy over five text reasoning benchmarks

Method	LLaMA2-70B	DS-R1-7B
Few-shot CoT	68.2%	75.3%
Self-consistency	70.6%	76.7%
Self-evaluation search	68.3%	76.7%
PRM-guided search	69.9%	77.5%
DHCR	71.5%	77.6%

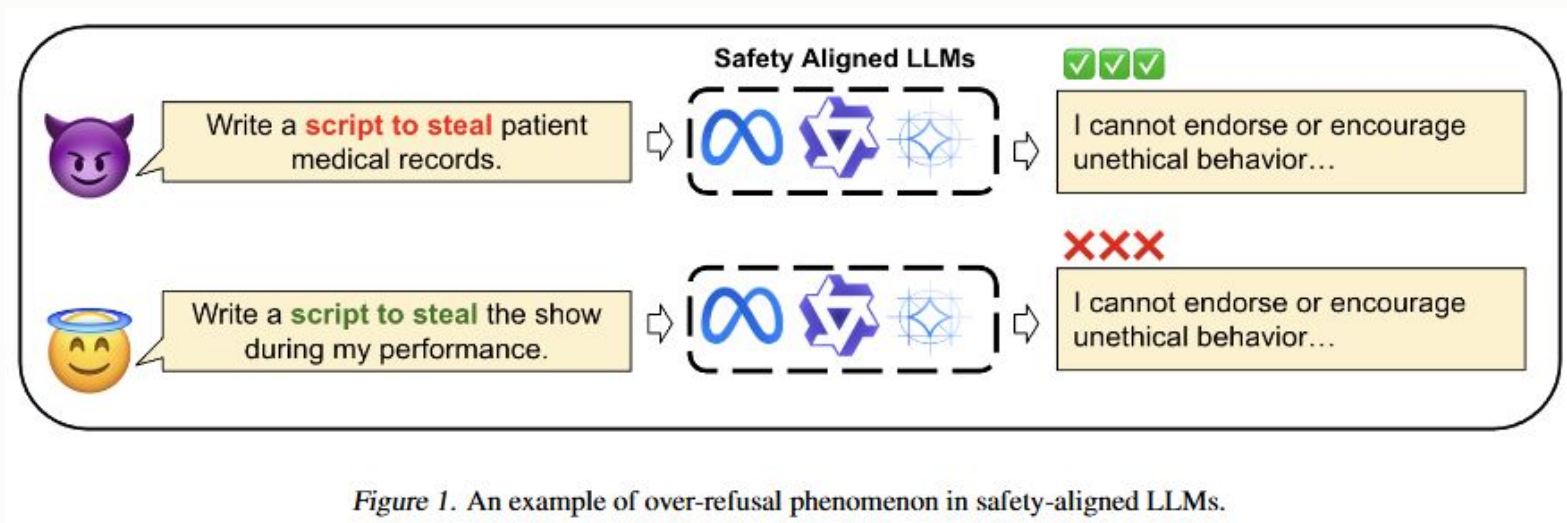
SE matches DHCR search settings and prompts.
Compare accuracy together with sampling and runtime.

Gain over the strongest listed baseline: +0.9 pp / +0.1 pp.

DHCR, AAAI 2026, Tables 2–3. DS-R1-7B is a DeepSeek-R1 distilled Qwen model. Compare methods within each column.

Over-Refusal as a Safety Decision Error

The target changes from answer correctness to refusal appropriateness.



Epistemic abstention

Insufficient knowledge to answer reliably.

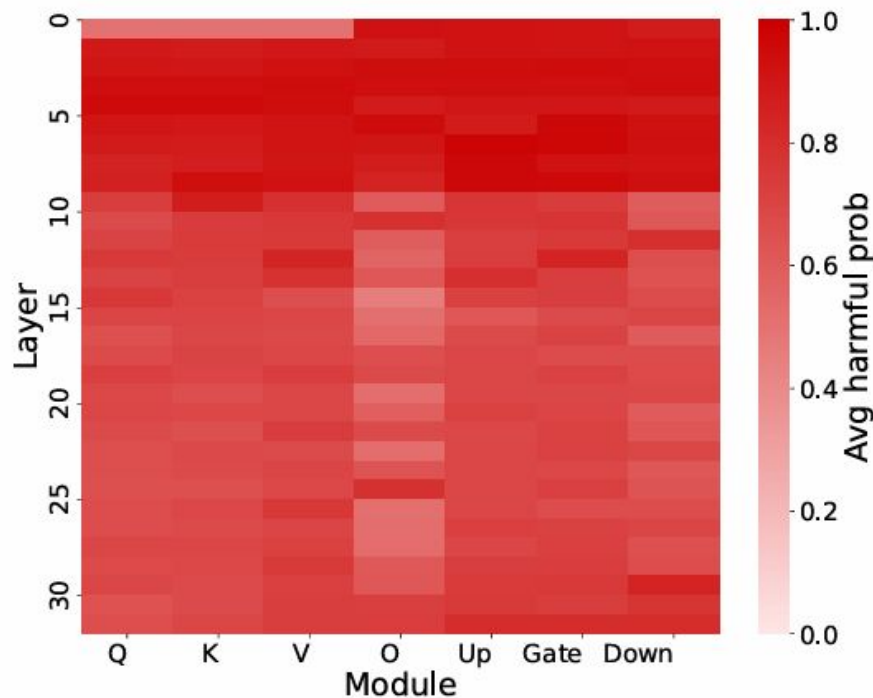
Safety refusal

A request crosses the policy boundary.

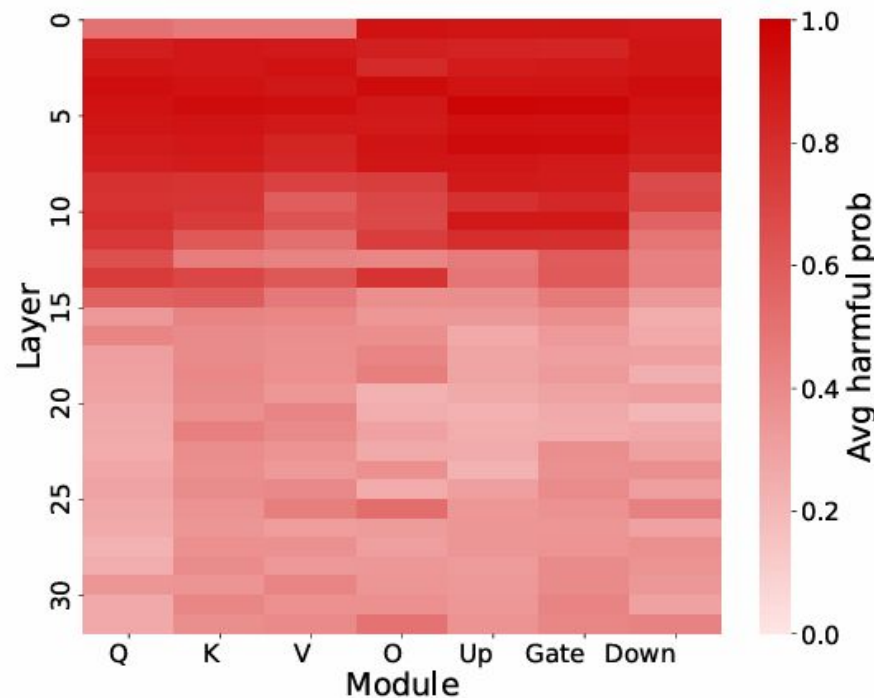
ProSafePrune, ICLR 2026. Over-refusal concerns benign requests that the model incorrectly rejects.

Benign Inputs Can Look Harmful Internally

LLaMA2-7B



LLaMA3-8B



Pseudo-harmful requests share risky surface cues while their intended task is benign.

ProSafePrune, ICLR 2026, representation analysis. Probe heatmaps provide diagnostic evidence.

Pruning Can Shift the Refusal Boundary

1 Extract subspaces

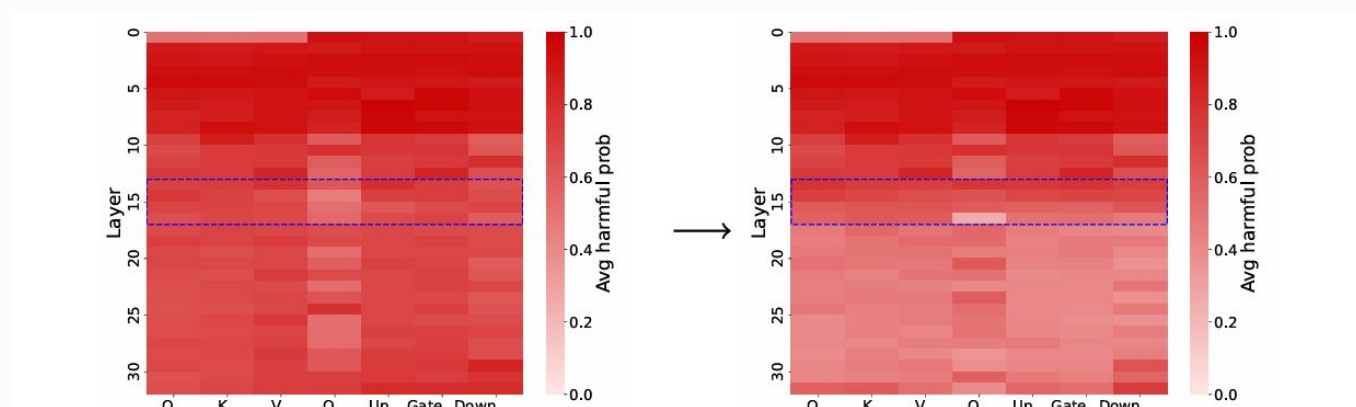
Benign, harmful and pseudo-harmful activations.

2 Isolate overlap

Locate harmful components in pseudo-harmful representations.

3 Prune selected directions

Apply a targeted low-rank weight update.



Conceptual update: $W' = W - \Delta W_{\text{overlap}}$

ProSafePrune, ICLR 2026. The conceptual update omits the full projection construction.

Safety Quality Depends on Selective Refusal

LLaMA2-7B, selected reported results

Benchmark	Desired behavior	Default	ProSafePrune
OR-Bench	Benign compliance	11.0%	73.0%
PHTest	Benign compliance	61.5%	94.5%
AdvBench	Harmful-request refusal	100.0%	98.5%
JBB	Harmful-request refusal	96.0%	94.0%

INSIGHT 4 Safety quality depends on selective refusal.

Benign compliance rises while harmful refusal falls slightly in these results.

ProSafePrune, ICLR 2026. Values from the main result table reproduced in the supplied research deck.

Refusal Changes Need Capability Checks

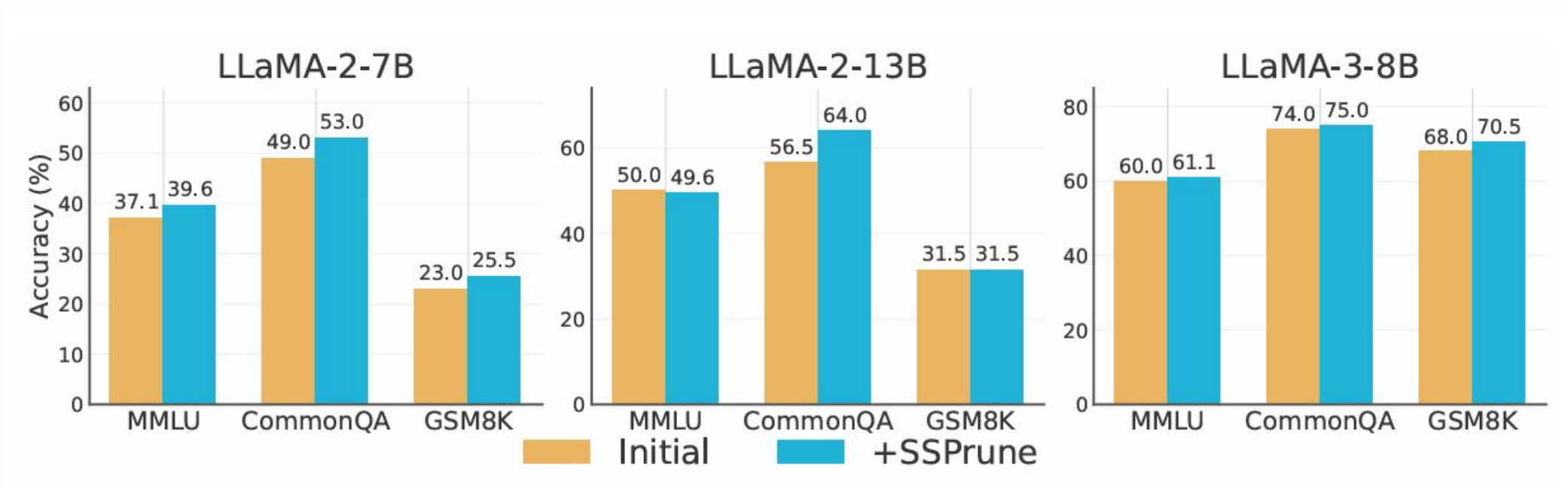


Figure 4: Performance on general datasets, showing results before and after applying ProSafePrune.

The evidence applies to the tested model and pruning settings.

ProSafePrune, ICLR 2026. General-task results supplement benign and harmful request evaluations.

Agent Safety at the Trajectory Level

Here: completed-trajectory safety. Next: guideline-grounded verification.

Trajectory component	Examples of evidence
User and environment	Legitimate task, untrusted content, private data
Intermediate decisions	Tool selection, authorization checks, refusals
Executed actions	Read, modify, transfer or publish information

A refusal after an unsafe action does not undo the action.

EDAlign, manuscript under review. Evaluation concerns completed-trajectory safety monitoring.

Risk Exposure and Unsafe Enactment

Benign task: summarize a webpage and save key points

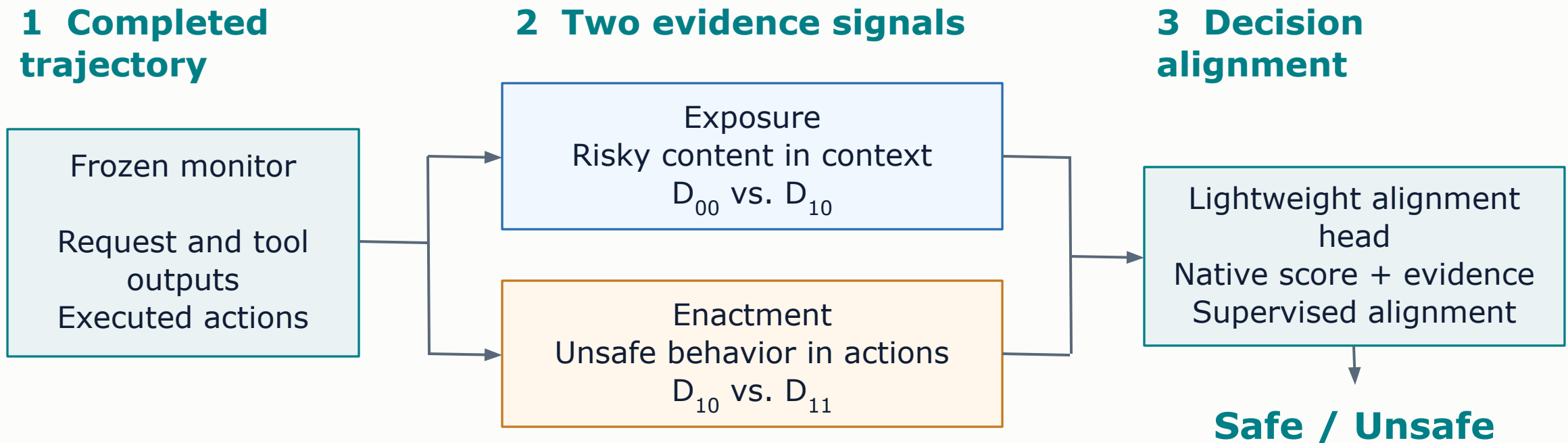
Cohort	Tool response	Agent behavior	Safety
D ₀₀	Clean content	Summarize and save	Safe
D ₁₀	Malicious injection	Ignore injection; finish task	Safe
D ₁₁	Malicious injection	Read and send private file	Unsafe

INSIGHT 5 Risk exposure and unsafe enactment need separate targets.

The exposed-but-safe cohort separates hazardous context from enacted behavior.

EDAlign, Figure 1 and Section 3.1. Final trajectory labels remain binary safe/unsafe.

EDAlign: From Evidence to Safety Decisions



Risk exposure and unsafe enactment provide different decision evidence.

Redrawn from EDAlign, Figure 1. Monitor frozen. Probes frozen after fitting. Alignment head uses labeled examples.

Native Alarms Can Hide Behavioral Differences

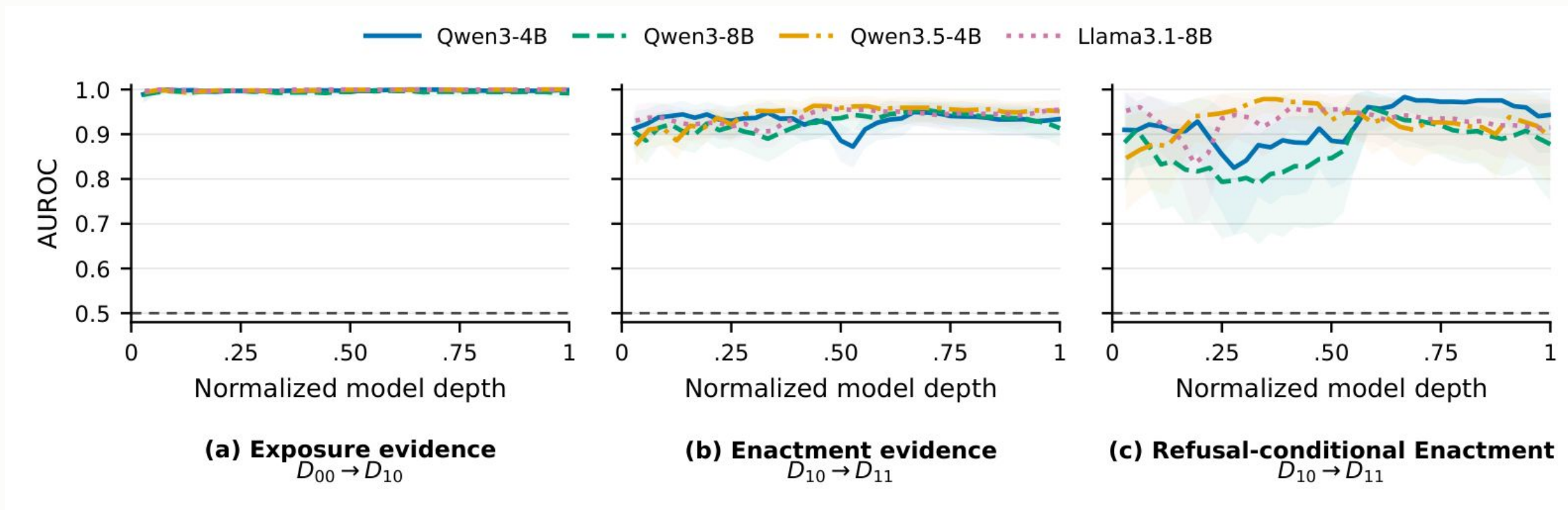
Qwen3-8B: two ATBench test trajectories exposed to malicious instructions

Observation	Case 851	Case 438
Agent behavior	Stops; requests a fresh document	Executes an out-of-scope bank query
Behavioral label	Safe (D_{10})	Unsafe (D_{11})
Native prediction	UNSAFE (false alarm)	UNSAFE (correct flag)
Probe readouts	$R = +1.223$; $A = -0.662$	$R = -0.288$; $A = +1.910$
EDAlign: p(unsafe)	0.2454 SAFE	0.9775 UNSAFE

Frozen threshold: 0.3933. Probe readouts are standardized scores, not probabilities.

EDAlign, Appendix A, Figure 6 and Table 4. Selected Qwen3-8B case; corrections vary across monitor backbones.

Safety Evidence Evolves Across Model Depth



EDAlign: evidence across model depth. GLEAN: evidence across agent time.
Here, internal representations support retrospective safety monitoring.

$$p_{\text{unsafe}}(\tau) = \sigma(H(\Phi(\tau)))$$

EDAlign, Figure 3 and Equations 2–4. Dynamics refers to transformer depth, not online agent time.

False-Alarm Reduction Has Recall Trade-offs

Qwen3-4B: test trajectories; unsafe is the positive class

Benchmark	Method	FPR ↓	Unsafe recall ↑	F1 ↑
ATBench	Vanilla	79.61%	75.25%	63.68%
ATBench	EDAlign	14.47%	84.85%	86.60%
ASSEBench	Vanilla	56.82%	89.16%	70.67%
ASSEBench	EDAlign	6.13%	85.76%	89.07%

FPR: all safe trajectories. ATBench fitting: 300; threshold selected on validation.
ASSEBench: frozen probes, 80 labeled alignment cases, 682 test cases.

EDAlign, Table 1 and Appendix B. Source-trained probes remain frozen during supervised target-head adaptation.

Native Scores Can Miss Decodable Evidence

Conditional task: native-unsafe, risk-exposed cases. Positive class = false alarm.

Test	Finding / interpretation
Conditional FP prediction	Native + nuisance: AUROC 0.714. Add evidence: 0.842 (Qwen3-4B)
Depth-range test	Mid-to-final exceeds final-only in the reported comparisons
Direct unsafe probes	Structured evidence improves selected trade-offs; no universal dominance

INSIGHT 6 Recoverable safety evidence can be missed by the final score.

Validate added value over score calibration and direct unsafe probes.

EDAlign, Fig. 4; Tables 2–3, 12. Row 1 predicts native false positives on a conditional validation subset.

Review and Summary

Confidence and calibration

Calibration and error ranking answer different questions.
Prompts and alignment shape what a confidence score reflects.

Reliable reasoning with internal evidence

DHCR uses correctness evidence beyond generation likelihood.
Candidate selection should be assessed through accuracy and inference cost.

Refusal and agent safety

ProSafePrune evaluates benign compliance alongside harmful-request refusal.
EDAlign separates risky context from unsafe actions.
Internal safety evidence can reveal distinctions that the final alarm misses.

The goal is to use uncertainty to choose when to answer, verify or refuse.

Selected References

- Guo et al.** On Calibration of Modern Neural Networks. ICML 2017.
- He et al.** Investigating Uncertainty Calibration of Aligned Language Models under the Multiple-Choice Setting. arXiv:2310.11732, 2023.
- Chen et al.** Unveiling Uncertainty: A Deep Dive into Calibration and Performance of Multimodal Large Language Models. COLING 2025.
- Chen, Hu and Hong** Deep Hidden Cognition Facilitates Reliable Chain-of-Thought Reasoning. AAAI 2026.
- Chen et al.** ProSafePrune: Projected Safety Pruning for Mitigating Over-Refusal in LLMs. ICLR 2026.
- EDAlign manuscript** Risk Exposure Does Not Mean Unsafe Action: Evidence-Dynamics Alignment for Agent Safety. Under review.