THINK WITHOUT THINKING: WHEN EMPTY REASONING REVEALS ANSWERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty before response generation would let a language model decide whether to answer, abstain, retrieve evidence, or spend more computation without first producing a candidate. Existing query-level observers usually read the final query token, implicitly treating observation position as fixed. We show that reasoning-model templates contain a more informative alternative: the first hidden state after the model's native empty-reasoning boundary. We call the resulting observer THINKUQ. It visits this boundary with a fixed teacher-forced suffix, decodes no tokens, and fits a lightweight correctness readout. On Qwen3-4B, THINKUQ improves subject-OOD MMLU AUROC by 9.95 points and transfers without target tuning to MMLU-Pro. Across four Qwen3 scales and tasks spanning factual recall, arithmetic, and code, the gain appears when task-relevant correctness is internally available before generation. Matched ordinary suffixes, official template variants, learned prefixes, activation patching, attention analysis, and a controlled Tagged-CoT training trajectory localize the effect to a learned post-boundary transition. In the positive regime, this transition widens correct–incorrect separation, shortens prequential codelength, and remains useful with few source labels and modest prefill overhead. These results establish observation position as a design variable for pre-answer uncertainty and provide a source-side diagnostic for recognizing when a native reasoning boundary exposes a useful correctness readout.

1 INTRODUCTION

A system that can estimate whether a language model will answer correctly *before* decoding can abstain, retrieve evidence, or allocate additional computation before committing to a response. This setting is stricter than scoring a completed answer: the observer cannot use answer likelihood, sampled rationales, agreement across generations, or post-hoc self-critique. Calibration and selective prediction provide the decision-theoretic foundation (Guo et al., 2017; Jiang et al., 2021; Kamath et al., 2020; Varshney et al., 2022); recent language-model methods estimate pre-answer confidence through verbal reports or query-side hidden states (Kadavath et al., 2022; Lin et al., 2022; Xiong et al., 2024; Chen et al., 2026; Gottesman & Geva, 2024; Wang et al., 2025; Moreno Cencerrado et al., 2026). A zero-decoding observer would preserve this information boundary while remaining cheap enough to guide the answer policy itself.

Pre-answer UQ has nevertheless treated one design choice as fixed: where the model is observed. Most methods read the final user-query token. Yet evidence about truthfulness, hallucination, semantic uncertainty, and answerability varies strongly across layers and token positions (Burns et al., 2023; Azaria & Mitchell, 2023; Chen et al., 2024; Kossen et al., 2024; Orgad et al., 2025; Slobodkin et al., 2023). The final query state may encode the relevant information without arranging it in a geometry that a small readout can access. Reasoning models create another natural observation point. Their native templates implement a learned transition from a reasoning phase to answer production (Yang et al., 2025). This raises a concrete question: can a model reveal future correctness more clearly after traversing that transition, even when the reasoning block is empty and no answer token is generated?

THINKUQ operationalizes this idea by appending the model's fixed empty-reasoning boundary and reading the first hidden state after it. The suffix is teacher-forced, query-independent, and carries no new correctness evidence; its role is to transform the representation through which existing query information is read. A standardized linear probe then estimates correctness at the post-boundary state.

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.

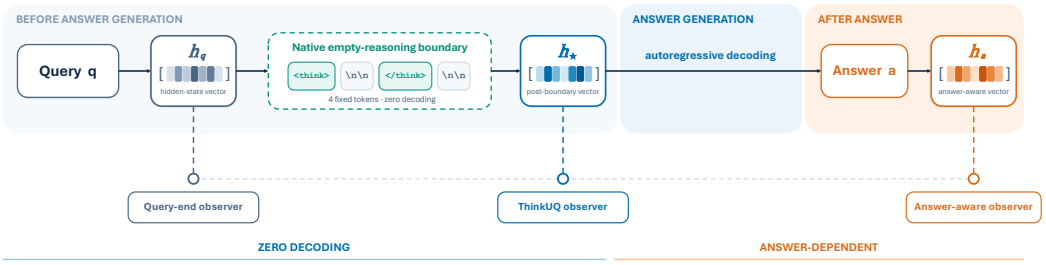

Figure 1: Observation positions for correctness UQ. Query-end reads the final query state h_q. THINKUQ teacher-forces the shown canonical four-token empty-reasoning boundary and reads the post-boundary state $h_\star$, decoding no reasoning or answer. Answer-aware methods read h_a after generating answer a.

This view makes observation position a controllable part of the UQ design rather than an incidental consequence of prompt termination.

The experiments reveal a recognizable positive regime. Under a 3,000/1,000 protocol, THINKUQ improves subject-OOD MMLU AUROC by 9.95 points and transfers without target tuning to MMLU-Pro with a 7.25-point gain. TriviaQA yields positive point estimates across four Qwen3 scales, while arithmetic and code map how the effect changes as correctness becomes more dependent on later computation. Token-position, ordinary-suffix, official-template, learned-prefix, and activation-patching controls localize the gain to the native post-boundary transition. Correctness-conditioned separation, fixed-layer margin, prequential codelength, attention routing, and a controlled Tagged-CoT training trajectory further show that the transition makes correctness information easier to read rather than merely changing global confidence.

This characterization also makes the effect usable. A grouped source-side diagnostic identifies whether the boundary increases both ranking quality and correct–incorrect separation before target evaluation. Once the observation state is selected and frozen, the gain remains accessible with 4–16 balanced source labels and a short fixed prefill, with no decoding. The method therefore connects a scientific finding about representation geometry to a practical pre-answer uncertainty observer.

Our contributions are:

- We introduce THINKUQ, a fixed-prefill, zero-decoding UQ observer that reads the first hidden state after a model's native empty-reasoning boundary.
- We identify a cross-task positive regime, including a $+9.95$-point subject-OOD MMLU gain and positive TriviaQA estimates across all four tested Qwen3 scales.
- We localize the gain to the learned post-boundary transition and connect it to increased correctness-conditioned separation, linear margin, prequential compression, attention routing, and Tagged-CoT exposure.
- We turn the finding into a source-side diagnostic and show that the selected readout remains useful with 4–16 balanced source labels, a short fixed prefill, and no decoded tokens.

2 RELATED WORK

Pre-answer confidence and selective prediction. Calibration relates confidence to empirical correctness, while selective prediction uses confidence to control coverage and error (Guo et al., 2017; Jiang et al., 2021; Kamath et al., 2020; Varshney et al., 2022). Language models express confidence through verbal probabilities, statements of knowledge, or $P(\text{IK})$ (Kadavath et al., 2022; Lin et al., 2022; Xiong et al., 2024). Query-level UQ estimates correctness before an answer is available (Chen et al., 2026). Hidden-state methods read entity knowledge or correctness from question representations (Gottesman & Geva, 2024; Wang et al., 2025; Moreno Cencerrado et al., 2026); DRIFT trains nonlinear detectors on intermediate question states (Bhatnagar et al., 2026). First-decoded-token representations can also expose answerability (Slobodkin et al., 2023). THINKUQ makes the observation state itself the intervention: it traverses a fixed native reasoning boundary

before decoding and compares states under matched correctness supervision and linear readouts. Query-difficulty controls distinguish this signal from global item difficulty (Seo et al., 2026).

Answer-aware and representation-based uncertainty. Answer likelihood, $P(\mathrm{True})$, linguistic confidence, self-checking, and agreement exploit evidence from generated candidates (Kadavath et al., 2022; Manakul et al., 2023; Xiong et al., 2024; Lin et al., 2024). Predictive entropy (Malinin & Gales, 2021), self-consistency (Wang et al., 2023), and semantic entropy over meaning-equivalent generations (Kuhn et al., 2023; Farquhar et al., 2024) likewise use answer-dependent evidence. Semantic Entropy Probes (SEPs) distill sampled-answer entropy into hidden-state predictors, including a token-before-generation readout (Kossen et al., 2024). Thus SEP and THINKUQ both support zero-decoding inference, while differing in supervision and observation state: SEP predicts entropy, whereas THINKUQ predicts correctness after a fixed native transition. Studies of latent truthfulness and hallucination further motivate low-complexity readouts (Burns et al., 2023; Azaria & Mitchell, 2023; Chen et al., 2024; Orgad et al., 2025). We isolate the contribution of the post-boundary state with matched linear probes and use answer-aware scores as richer-information references.

Reasoning transitions and latent computation. Chain-of-thought prompting, zero-shot reasoning, self-consistency, and rationale fine-tuning show that intermediate computation and its serialization can change model behavior (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023; Zelikman et al., 2022; Kim et al., 2023). Learned pause tokens likewise provide useful pre-output computation (Goyal et al., 2024). Meaningless fillers can supply hidden computational depth when models are densely trained to use them on algorithmic tasks (Pfau et al., 2024); at inference time, filler type, count, and location can also improve QA and mathematics accuracy (Jang et al., 2025). These studies ask whether added token positions improve answer computation. THINKUQ fixes the answer policy and uses a short template-native suffix to expose a zero-decoding correctness readout. Generated-question interventions provide a richer diagnostic state (Luo et al., 2026), whereas THINKUQ generates neither question nor rationale. Position, suffix, official-template, soft-prefix, patching, and training-trajectory controls test how the post-boundary transition makes query-induced correctness information easier to read.

3 PRE-ANSWER UQ AT A NATIVE BOUNDARY

3.1 PROBLEM AND INFORMATION BOUNDARY

Let f_θ be a frozen autoregressive model, q a query, and π a fixed answer policy. The offline answer pass produces $a \sim \pi(f_\theta, q)$ and label

$$y(q) = \mathbb{I}\{a \in \mathcal{A}_{\mathrm{correct}}(q)\}. \tag{1}$$

The observer estimates $p(y = 1 \mid q, f_\theta, \pi)$ before decoding any token from a. The answer and observer passes are strictly separated: the answer pass always receives the original prompt, while a diagnostic prefill is introduced only to the observer. Hence the method predicts correctness of the unchanged answer policy rather than improving or perturbing that answer.

Query-end observers read f_θ at the final state of the original query. Our setting additionally permits a fixed, answer-independent suffix, but forbids model-generated tokens, selected answer choices, rationales, retrieval outcomes, and correctness feedback. Answer-aware methods may condition on those post-generation quantities. This places THINKUQ on the query side of the information boundary even though its readout is located in the assistant turn.

3.2 EMPTY REASONING AS A STATE-ELICITING INTERVENTION

Let $\tau(q)$ be the official chat serialization through the beginning of the assistant turn, and let $b_\emptyset$ be the model's fixed empty/disabled-reasoning suffix. The primary MMLU protocol inserts a five-token empty block before a shared one-token newline readout used in every condition. Canonical-template panels use Qwen3's official four-token empty suffix directly. Appendix G gives both token sequences. Every suffix is teacher-forced and query-independent. We read

$$\boldsymbol{z}_\ell^{\mathrm{empty}}(q) = h_{\ell,t_\star}(\tau(q) \oplus b_\emptyset)\,, \tag{2}$$

where $t_\star$ is the post-boundary readout position, the shared newline in the primary protocol. The Query-end baseline extracts $\boldsymbol{z}_\ell^{\mathrm{query}}(q)$ at the final state of the original query. For each candidate layer we fit the same standardized ℓ_2-regularized logistic readout,

$$\widehat{p}_\ell(y = 1 \mid q) = \sigma\big(\boldsymbol{w}_\ell^\top \operatorname{Std}(\boldsymbol{z}_\ell(q)) + b_\ell\big)\,. \tag{3}$$

Layer and regularization are selected using source validation NLL. The standardizer, probe, layer, and calibration are then frozen for source test and target transfer. Shared- and globally fixed-layer analyses measure within-layer accessibility (Appendix A.6); the training trajectory also keeps both observers at layer 30.

Because $b_\emptyset$ is fixed and independent of (q, y), it cannot add exogenous correctness information:

$$I(y; q, b_\emptyset) = I(y; q). \tag{4}$$

The boundary changes the representation through which that information is read. If $\mathcal{R}_{\mathrm{lin}}(H)$ is the optimal logistic readout risk, deterministic representations of the same query can have different $\mathcal{R}_{\mathrm{lin}}$. The post-boundary transition can reorganize query-conditioned information into geometry that is more accessible to a small readout. This motivates tests of token position, equal-length suffixes, template semantics, and correct–incorrect separation. Appendix C relates this geometry to ranking and readout risk.

This prediction is stronger than a generic hidden-state change. A query-independent translation, $\boldsymbol{z}^{\mathrm{empty}} = \boldsymbol{z}^{\mathrm{query}} + \boldsymbol{c}$, is absorbed by the probe intercept without changing ranking. The positive regime therefore requires a query-dependent transformation that changes correctness geometry, motivating the matched-suffix controls and class-conditional diagnostics below.

The relevant geometric quantity is separation along a readable direction relative to within-class variation. In a shared-covariance Gaussian model, the best affine-score AUROC increases with the Mahalanobis distance between correct and incorrect class means (Appendix C). This gives the geometry analysis a concrete target: the boundary should improve class separation, not merely move all states farther from the query anchor. The fixed-layer margin and source-frozen transfer tests examine that prediction from complementary directions.

4 Experimental Design

The primary evaluation uses Qwen3-4B (Yang et al., 2025) on MMLU subject-OOD and MMLU→MMLU-Pro transfer (Hendrycks et al., 2021a; Wang et al., 2024): 3,000 source and 1,000 target examples with ten frozen source-subject folds. The cross-scale study evaluates Qwen3-1.7B, 4B, 8B, and 14B. Tasks cover mathematics (Cobbe et al., 2021; Hendrycks et al., 2021b), factual QA (Joshi et al., 2017; Mallen et al., 2023), and code (Austin et al., 2021; Chen et al., 2021), with additional science and expert QA references (Clark et al., 2018; Rein et al., 2024). Free-form and code correctness use frozen normalized or execution-based evaluators, separately from observer inputs.

We report answer accuracy and AUROC as the primary discrimination measure, with AURC, NLL, Brier score, ECE, and fixed-coverage risk reported separately. Unless noted otherwise, transfer inference uses 5,000-draw paired hierarchical percentile bootstrap. "Ex" resamples frozen source splits and correctness-stratified target examples; "Gp" resamples the same splits and target semantic groups. We call a direction resolved only when both 95% intervals exclude zero, emphasizing stability across examples and semantic groups. For TriviaQA, we additionally average the shared-source-layer deltas over all four tested Qwen3 scales and jointly resample aligned source splits and target examples or groups. The intervals quantify sampling variation under fixed models, answer policies, and recorded seeds; they are not multi-seed estimates. Appendix F gives prompts, answer policies, data splits, and implementation details.

5 When Does Empty Reasoning Reveal Answerability?

5.1 ThinkUQ outperforms matched answer-free UQ methods

We begin with the principal method comparison. On subject-disjoint MMLU, the Qwen3-4B Query-end observer reaches .6900 AUROC, while ThinkUQ reaches .7895. The +.0995 gain has paired

Table 1: Qwen3-4B correctness UQ on MMLU subject-OOD and zero-tuning MMLU→MMLU-Pro (3,000/1,000 source/target examples). Answer-free probes use only MMLU source subjects; baselines are ordered by publication year, with years denoting the cited method or closest implementation family. No Answer Needed uses DoM and Internal Confidence uses grid LR. Bold marks the best answer-free method per column; Choice Summary is answer-aware. Arrows show the preferred direction.

	MMLU subject-OOD			MMLU→MMLU-Pro		
Method	AUROC↑	AURC↓	NLL↓	AUROC↑	AURC↓	NLL↓
Class prior	.5000	.2644	.5931	.5000	.5633	.8660
Question TF–IDF	.5879	.2295	.5870	.5898	.5087	.8370
Query-only (flat) (2023)	.6335	.1990	.5793	.5859	.4956	.7529
Matched filler (2024; 2025)	.7264	.1419	.5305	.6588	.4237	.6921
FacLens (2025)	.6676	.1743	.5735	.6165	.4709	.7094
Query-end state (LR) (2026)	.6900	.1595	.5544	.6524	.4359	.6905
No Answer Needed (DoM) (2026)	.6077	.2127	.5783	.5917	.5044	.7581
Internal Confidence (grid LR) (2026)	.6053	.2187	.5900	.5910	.4985	.7580
THINKUQ (ours: post-boundary state)	**.7895**	**.1153**	**.4828**	**.7249**	**.3696**	**.6750**
Choice Summary (answer-aware) (2022)	.8044	.1101	.4677	.7392	.3571	.6060

fold-bootstrap CI $[+.0802, +.1189]$. The same source-trained observers transfer without tuning from MMLU to MMLU-Pro at $.6524/.7249$, a $+.0725$ gain with split+example CI $[+.0441, +.0999]$ and split+subject CI $[+.0511, +.0929]$.

Table 1 compares UQ methods under the same 3,000/1,000 protocol. THINKUQ leads the compared answer-free methods on AUROC, AURC, and NLL for both subject-disjoint evaluation and dataset transfer. It improves over text-only difficulty predictors, Query-end representations, an equal-length matched filler, and matched versions of FacLens, No Answer Needed, and Internal Confidence. Choice Summary uses answer-choice information and provides an answer-aware reference. The three metrics capture ranking, selective prediction, and probabilistic loss, respectively.

5.2 WHEN IS CORRECTNESS AVAILABLE BEFORE DECODING?

Table 2 asks when task-relevant correctness information becomes available before decoding. Throughout this table, Query-end state denotes the same standardized LR baseline used in Table 1. The three targets span increasingly trajectory-dependent correctness. TriviaQA tests whether retrieved factual knowledge already carries a known-versus-unknown signal. GSM8K tests whether a latent multi-step solution has been organized before it is verbalized. HumanEval tests correctness that depends on a realized executable program.

TriviaQA question-source transfer provides the clearest cross-scale positive regime. On the expanded 1,200-example target, every tested Qwen3 scale has a positive point estimate; the 4B and 14B gains are resolved under both example and source-group bootstrap, while the 1.7B and 8B estimates remain unresolved. An aligned 600-example shared-layer synthesis averages the four scale-specific deltas before inference and yields a $+2.82$-point family effect, with example CI $[+1.13, +4.55]$ and group CI $[+1.76, +3.68]$. Condition-specific and globally fixed-layer sensitivities of that family statistic are also positive under both units (Appendix A.6). Factual answerability is therefore visible at the boundary across scales, with the clearest individual gains at 4B and 14B.

The pattern changes as correctness depends on more computation after the boundary. On the full 1,319-example GSM8K test set, all four point estimates are positive. The 8B gain is resolved under both resampling units; the 14B example interval narrowly includes zero, and the 1.7B and 4B comparisons remain unresolved. This scale-selective evidence is consistent with the post-boundary state becoming useful when a checkpoint has consolidated enough of the solution structure before emitting the answer. All four HumanEval comparisons are unresolved. A plausible explanation is that program correctness depends more strongly on the complete multi-token program and its execution semantics, limiting how much of the final outcome a single pre-decoding state can expose. Together, the three tasks identify a representation-readiness condition rather than a uniform task or scale effect: THINKUQ helps when correctness has become internally available before generation.

Table 2: Cross-scale Qwen3 correctness UQ on TriviaQA, GSM8K, and HumanEval. Acc. is answer accuracy; Query-end and THINKUQ are AUROCs; Gain is THINKUQ minus Query-end. Ex/Gp are paired 95% split+target-example/group bootstrap CIs; bold gains exclude zero under both units.

Params.	Transfer target	Acc.	Query-end state	THINKUQ	Gain	Ex 95% CI	Gp 95% CI
1.7B	TriviaQA	.148	.7490	.7629	$+.0139$	$[-.0184, +.0464]$	$[-.0261, +.0432]$
4B	TriviaQA	.241	.7800	.8330	$+.0530$	$[+.0271, +.0785]$	$[+.0249, +.0764]$
8B	TriviaQA	.391	.7897	.8030	$+.0133$	$[-.0028, +.0299]$	$[-.0049, +.0249]$
14B	TriviaQA	.427	.7803	.8162	$+.0359$	$[+.0190, +.0526]$	$[+.0170, +.0590]$
1.7B	GSM8K	.796	.6481	.6518	$+.0037$	$[-.0347, +.0397]$	$[-.0286, +.0329]$
4B	GSM8K	.890	.5655	.5711	$+.0057$	$[-.0324, +.0449]$	$[-.0108, +.0211]$
8B	GSM8K	.313	.7091	.7680	$+.0589$	$[+.0349, +.0849]$	$[+.0441, +.0778]$
14B	GSM8K	.346	.8162	.8280	$+.0118$	$[-.0013, +.0249]$	$[+.0082, +.0162]$
1.7B	HumanEval	.573	.5586	.5084	$-.0501$	$[-.1308, +.0283]$	$[-.1295, +.0290]$
4B	HumanEval	.695	.4619	.5187	$+.0568$	$[-.0150, +.1339]$	$[-.0160, +.1325]$
8B	HumanEval	.817	.5914	.6067	$+.0152$	$[-.0647, +.1002]$	$[-.0697, +.1047]$
14B	HumanEval	.872	.5368	.5705	$+.0338$	$[-.0647, +.1371]$	$[-.0668, +.1320]$

Table 3: Native-boundary controls. The ordinary-suffix test uses Qwen3-14B TriviaQA; others use Qwen3-4B MMLU→MMLU-Pro. Rows give subtraction order. Ex/Gp are paired 95% target-example/task-group (question-source or subject) CIs; all exclude zero.

Contrast	Δ AUROC	Ex 95% CI	Gp 95% CI
THINKUQ − 20 ordinary-suffix mean	$+.0837$	$[+.0433, +.1274]$	$[+.0354, +.1243]$
Official no-thinking − generation boundary	$+.0928$	$[+.0257, +.1648]$	$[+.0168, +.1739]$
Learned soft prefix − THINKUQ	$-.1662$	$[-.2422, -.0969]$	$[-.2338, -.1063]$
Query-end patch − identity THINKUQ pass	$-.1729$	$[-.2732, -.0691]$	$[-.2578, -.0962]$

5.3 NATIVE-BOUNDARY CONTROLS LOCATE THE READOUT GAIN

The gain is localized in both position and construction. In the primary token-position control, the exact opening and closing tag states are weak; the best state is the first shared token *after* the complete boundary, and later fixed readout tokens decline (Appendix B). Structure also matters. On a fixed Qwen3-14B TriviaQA control, THINKUQ exceeds the mean of twenty randomly drawn four-token ordinary suffixes by $+.0837$ AUROC; none of the twenty reaches its point estimate. Rendering the checkpoint's official no-thinking mode reproduces the THINKUQ prefill token-for-token and improves over the assistant-generation boundary by $+.0928$.

A source-learned four-vector soft prefix, initialized from the native boundary embeddings, returns to Query-end-level target performance and trails THINKUQ by $.1662$ AUROC. In a matched Qwen3-4B MMLU-Pro sample, replacing the THINKUQ residual stream with its paired Query-end state reduces AUROC by $.1729$. Table 3 combines these tests with the suffix and template controls. The native transition outperforms matched ordinary tokens and learned prefix capacity, reproduces the official template, and supplies a residual state whose replacement disrupts the readout. These results connect the gain to the computation elicited by the native boundary.

5.4 WHY THE POSITIVE REGIME WORKS: CORRECTNESS BECOMES EASIER TO READ

AUROC records ordering, but not why the ordering changes. We therefore test the Qwen3-14B TriviaQA positive regime with four independently defined diagnostics (Table 4). The boundary raises the mean predicted probability on correct answers by $.0270$ and lowers it on incorrect answers by $.0255$, increasing their separation by $+.0525$. At a globally fixed 75%-depth layer, correct-minus-incorrect logit separation grows by $.1855$ pooled standard deviations. A source-only prequential code also shortens by $.4401$ bits/example, without using target states or labels. The target ranking gain, target class separation, fixed-layer margin, and source codelength therefore agree: the native transition makes future correctness easier for a linear observer to read.

The joint ranking and separation gains connect THINKUQ to a concrete geometric account: the native transition exposes a stronger correct–incorrect contrast to the linear readout. Appendix C derives

Table 4: Mechanistic diagnostics on 600 Qwen3-14B TriviaQA targets. Separation is the correct-minus-incorrect probability gap; margin is the standardized logit gap; codelength is source bits/example. Effects are THINKUQ minus Query-end (negative favors codelength). Ex/Gp are paired 95% example/question-source CIs; all effects are favorable under both units.

Diagnostic	Query-end	THINKUQ	Effect	Ex 95% CI	Gp 95% CI
Target AUROC	.7664	.7993	$+.0330$	$[+.0122, +.0545]$	$[+.0020, +.0581]$
Probability separation	–	–	$+.0525$	$[+.0310, +.0733]$	$[+.0215, +.0754]$
Fixed-layer margin	.9958	1.1813	$+.1855$	$[+.0668, +.3017]$	$[+.0341, +.3161]$
Prequential bits/example	2.0823	1.6423	$-.4401$	$[-.5957, -.2787]$	$[-.7012, -.2045]$

Table 5: Qwen3-14B TriviaQA few-label UQ after frozen source selection. Budgets are balanced source labels/split; Full uses all labels. Query-end/THINKUQ are AUROC on 600 targets; Gain is THINKUQ minus Query-end. Ex/Gp are paired 95% target-example/question-source CIs; all gains exclude zero.

Labels/split	Query-end	THINKUQ	Gain	Ex 95% CI	Gp 95% CI
4	.5133	.5534	$+.0401$	$[+.0075, +.0748]$	$[+.0010, +.0718]$
8	.5352	.6228	$+.0876$	$[+.0573, +.1217]$	$[+.0613, +.1143]$
16	.5586	.6343	$+.0758$	$[+.0483, +.1046]$	$[+.0471, +.1004]$
Full	.7075	.7514	$+.0439$	$[+.0064, +.0805]$	$[+.0129, +.0821]$

the relationship between score separation and AUROC, while Appendix A.8 reports probability calibration and selective risk. The layer, geometry, and prompt-surface controls further characterize this change in accessibility.

6 WHAT CAN THE POSITIVE REGIME BUY?

6.1 THE GAIN SURVIVES VERY SMALL LABEL BUDGETS

The readout is supervised, so we test label efficiency after the grouped source-side diagnostic has selected a configuration. We freeze the layer, regularization, ten source splits, and seed, then fit balanced Qwen3-14B TriviaQA readouts with 4, 8, or 16 source labels per split. THINKUQ wins at every fit budget under both paired units on the fixed 600-example target (Table 5). The gain remains positive with only four labels and is largest at the 8- and 16-label settings. A diagonal-Gaussian typicality score without correctness labels does not reproduce the gain (Appendix A).

6.2 ZERO DECODING PRESERVES A DISTINCT COST–INFORMATION POINT

On a fixed 200-answer Qwen3-4B TriviaQA sample, the Query-end state reaches .7943 AUROC and the THINKUQ post-boundary state .8238 (Table 6). Sequence likelihood reaches .8641 after the answer has been generated; the answer-aware reference set observes a mean of 7.96 visible answer tokens. First-token and length-normalized likelihood do not exceed THINKUQ. Thus THINKUQ occupies a useful point between a bare query representation and post-answer evidence.

The incremental cost of reaching that point is small and explicit. On a matched-hardware Qwen3-14B TriviaQA run with 64 paired target queries, mean latency is 107.18 ms for Query-end and 115.00 ms for THINKUQ prefill: $+7.82$ ms, or $+7.30\%$. The paired example CI is $[4.64, 11.27]$ ms and the seven-group CI is $[0.90, 13.00]$ ms. In this configuration, the canonical suffix adds four input tokens, 10.65% analytic transformer FLOPs, .56 MiB mean incremental allocation, and zero decoded tokens. The observer therefore obtains the additional readout with a single short prefill.

6.3 RECOGNIZING THE POSITIVE REGIME FROM SOURCE DATA

The experiments support a source-side diagnostic. For a fixed model, answer policy, and task, compare Query-end and THINKUQ on grouped labeled source data and require a paired ranking gain accompanied by increased correct–wrong separation. The layer and readout are then frozen before

Table 6: Qwen3-4B TriviaQA information–cost comparison on a fixed 200-example sample (mean 7.96 answer tokens). Query-end/THINKUQ decode none; likelihood scores require generated answers. Values are descriptive AUROC on the same sample; bold marks the best zero-decoding score.

Score	Information available	Decoded for UQ	AUROC
Query-end state	Original query	0	.7943
THINKUQ post-boundary state	Query + 4 fixed tokens	0	**.8238**
First-token likelihood	Generated answer	Answer-dependent	.7155
Length-normalized likelihood	Generated answer	Answer-dependent	.7931
Sequence likelihood	Generated answer	Answer-dependent	.8641

target evaluation. In the positive TriviaQA regime, fitting the final readout remains effective with 4–16 balanced labels per split. Applications that require calibrated probabilities can calibrate the frozen score on separate outcomes; Appendix A.8 reports calibration and selective-risk contrasts.

6.4 THE ADVANTAGE SURVIVES INSTRUCTION PARAPHRASING

A useful readout should not depend on one exact wording of the prompt. We therefore run a matched Qwen3-4B MMLU→MMLU-Pro surface test with 400 source and 200 target examples. On the canonical prompt, Query-end/THINKUQ AUROC is $.6061/.6867$, a $+.0807$ gain with paired fold+example CI $[+.0196, +.1414]$ and fold+group CI $[+.0012, +.1615]$. A deterministic paraphrase changes the instruction and field labels while preserving every question, answer option, and target label. Under independently matched fitting, Query-end/THINKUQ becomes $.5772/.6737$: the $+.0965$ gain remains positive under both units (CIs $[+.0249, +.1735]$/$[+.0085, +.1920]$).

The positive regime is therefore not tied to the literal canonical instruction: it persists when the surrounding surface form changes while the task and native boundary remain intact. A paired difference-in-differences test does not detect a change in the gain between canonical and paraphrased prompts. Appendix Table 19 reports the complete matched and zero-adaptation surface panel.

7 HOW A POSITIVE READOUT STATE EMERGES DURING TRAINING

Starting from the official non-thinking Qwen3-4B-Instruct-2507 checkpoint, we follow a controlled Tagged-CoT continuation for 600 minutes. Every 30 minutes, the same 1,200-example MMLU source pool fits both observers and reads them at a prespecified common layer 30. Figure 2 shows MMLU→MMLU-Pro ΔAUROC (THINKUQ minus Query-end) on a fixed 600-example target at all 21 checkpoints.

At initialization, the example-bootstrap interval overlaps zero ($+.0080$, 95% CI $[-.0226, +.0420]$). The relative advantage strengthens with training time (Spearman $\rho = .660$, $p = .0011$): all 21 point estimates are positive, 16 example-bootstrap intervals exclude zero, and the 600-minute checkpoint reaches $\Delta = +.1160$ with CI $[+.0557, +.1774]$ ($p = .0010$). The common-layer design rules out checkpoint-wise layer relocation, supporting the account that Tagged-CoT exposure progressively makes correctness more readable at the post-boundary state relative to the query boundary. Appendix D.1 gives protocol details.

8 DISCUSSION

What the model appears to reveal. Position, suffix, template, learned-prefix, and activation-patching controls support a learned transition that consolidates answer-preparation signals. Tagged-CoT strengthens relative readability, and cross-layer routing shows that the representation is distributed.

Why the positive regime is scientifically useful. Applicability follows representation readiness: TriviaQA is positive at every scale and resolves at 4B/14B; GSM8K is positive throughout and resolves most clearly at 8B; HumanEval marks the trajectory-dependent boundary. On TriviaQA, low-label fitting agrees with class-conditional shifts, margin, and prequential codelength.

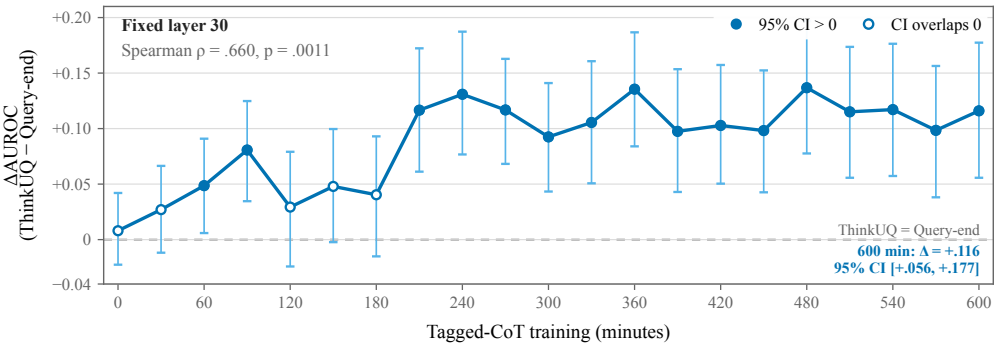

Figure 2: Tagged-CoT emergence in Qwen3-4B-Instruct-2507 over 21 checkpoints/600 minutes, with fixed layer 30 and MMLU→MMLU-Pro pools (1,200/600 source/target). Points are ΔAUROC (THINKUQ minus Query-end); bars are 2,000-draw paired example-bootstrap 95% CIs. Filled markers denote CIs above zero; open markers denote CIs overlapping zero; the dashed line marks zero.

The evidence chain. MMLU and MMLU-Pro establish the answer-free gain; cross-task and paraphrase panels map its availability. Native-suffix, geometry, compression, and trajectory analyses connect the gain to a learned transition, while transfer and low-label evidence show that the state is readable.

How the transition routes evidence. Attention changes concentrate in layers 20–31, recurring at layers 24, 26, 28, and 31. At layer 26, the post-boundary state assigns $+.059$ more mass to the full question and less to the recent 32–128-token tail; mean distance, maximum weight, and concentration rise by $+.091$, $+.016$, and $+.026$. These changes are stronger on correct examples. Together with activation patching, the pattern supports a transition that retrieves distributed query evidence into a more readable state (Appendix D.3).

Deployment perspective. THINKUQ supplies a correctness score to white-box systems before answer generation. Source correctness labels fit the observer, and the native boundary supplies a low-cost readout location. Once fitted, the observer leaves the backbone and answer policy unchanged. Applications can use the score for selective prediction or calibrate it to a probability scale using separate labeled outcomes. The Qwen3 results demonstrate how a native transition can support this design across model scales without adding a decoding stage.

9 CONCLUSION

THINKUQ makes observation position a design variable for pre-answer uncertainty. Teacher-forcing a model's native empty-reasoning boundary exposes a zero-decoding correctness signal more readable than the final query state. Cross-task results and controlled interventions trace this gain to a learned transition that reorganizes query evidence. This identifies model-native control transitions as useful observation sites for latent correctness. The geometric account connects this transition to class separation along readable directions while holding the probe family fixed. Ranking, margin, and codelength then characterize complementary benefits of the resulting representation. The training trajectory complements this static geometry: a growing relative advantage is visible within one fixed layer during Tagged-CoT continuation.

A grouped source-side diagnostic makes this observation actionable by selecting a state whose ranking and correct–incorrect separation improve together. The resulting label-efficient, fixed-prefill observer supplies a score for abstention, retrieval, or further computation without generating a candidate. This separates the cost of estimating correctness from the cost of producing an answer and creates an inexpensive interface between model representations and downstream decisions. Readout position and scoring function should therefore be designed jointly for pre-answer UQ.

Reproducibility Statement

All experiments use fixed seeds and source-only model selection, with answer generation separated from observer computation. The appendix specifies tokenization, prompts, data splits, evaluators, and bootstrap units. The code accompanying this work is included in the anonymous submission.

AI Use Statement

Generative AI tools were used in the preparation of this work to assist with language editing and polishing, the generation and debugging of experimental code, and the statistical summarization of experimental results. All AI-assisted outputs were carefully reviewed, verified, and, where necessary, revised by the authors. The authors take full responsibility for the content, methodology, experimental results, and conclusions presented in this paper.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

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

Rohan Bhatnagar, Youran Sun, Chi Andrew Zhang, Yixin Wen, and Haizhao Yang. DRIFT: Detecting representational inconsistencies for factual truthfulness. *arXiv preprint arXiv:2601.14210*, 2026. doi: 10.48550/arXiv.2601.14210. URL https://arxiv.org/abs/2601.14210v2.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0d1986a61e30e5fa408c81216a616e20-Abstract-Conference.html.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models. In *International Conference on Learning Representations*, volume 2026, pp. 34353–34372, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a07c3a67cfe50d3236b71fb674c7f30-Abstract-Conference.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

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. URL https://arxiv.org/abs/2110.14168.

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

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.232. URL https://aclanthology.org/2024.emnlp-main.232/.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/76917808731dae9e6d62c2a7a6afb542-Abstract-Conference.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Yoonna Jang, Kisu Yang, and Isabelle Augenstein. Expanding computation spaces of LLMs at inference time. *arXiv preprint arXiv:2509.24884*, 2025. doi: 10.48550/arXiv.2509.24884. URL https://arxiv.org/abs/2509.24884.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

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. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12685–12708. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.782. URL https://aclanthology.org/2023.emnlp-main.782/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://icml.cc/virtual/2024/39442.

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. URL https://openreview.net/forum?id=tWS-S_aRDRe.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=DWkJCSxKU5.

Chu Fei Luo, Samuel Dahan, and Xiaodan Zhu. What am i missing? question-answering as hidden state probing. *arXiv preprint arXiv:2605.31561*, 2026. URL https://arxiv.org/abs/2605.31561.

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

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*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

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

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 Trustworthy AI*, 2026. URL https://openreview.net/forum?id=OhN25uxVab.

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*, volume 2025, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=NikbrdtYvG.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Yeongbin Seo, Dongha Lee, and Jinyoung Yeo. Quantifying genuine awareness in hallucination prediction beyond question-side shortcuts. *arXiv preprint arXiv:2509.15339*, 2026. URL https://arxiv.org/abs/2509.15339.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.220. URL https://aclanthology.org/2023.emnlp-main.220/.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Towards improving selective prediction ability of NLP systems. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 221–226. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.repl4nlp-1.23. URL https://aclanthology.org/2022.repl4nlp-1.23/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yanling Wang, Haoyang Li, Hao Zou, Jing Zhang, Xinlei He, Qi Li, and Ke Xu. FacLens: Transferable probe for foreseeing non-factuality in fact-seeking question answering of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18563–18582. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.937. URL https://aclanthology.org/2025.emnlp-main.937/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6733cf15e10e2cd1d59af033c3bb8507-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

A Complete Experimental Results

This appendix expands the evidence for the method, its positive regime, state specificity, training-time emergence, robustness, and computational cost. It also provides the exact information boundary, source-only selection protocol, and paired statistical analyses used throughout the paper.

A.1 ThinkUQ exposes a stronger pre-generation signal

Table 7 reports the primary source-to-target protocol. ThinkUQ improves over the matched Query-end probe by 9.95 AUROC points on subject-OOD MMLU and 7.25 points on MMLU-Pro. All ten frozen folds favor ThinkUQ in both domains. On transfer, paired hierarchical intervals exclude zero under both example and subject resampling. ThinkUQ also exceeds the length-matched filler by 6.61 points, again under both units.

The source paired delta is $+0.0995$ with paired frozen-fold CI $[+0.0802, +0.1189]$. On target, the corresponding fold CI is $[+0.0659, +0.0809]$. Answer-conditioned confidence provides a richer-information reference for this zero-decoding comparison.

A.2 Native tags strengthen transfer

A structure control removes the reasoning tags while preserving the fixed newline layout. It reaches 0.7949 on MMLU, slightly above ThinkUQ (0.7886 in the matched boundary-control setting), but transfers worse to MMLU-Pro: 0.7130 versus 0.7254. The paired target difference is $+0.0124$ with 95% CI $[+0.0107, +0.0140]$, positive for all ten partitions. Correctly ordered tags also outperform swapped tags by $+0.0050$ in source and $+0.0101$ in target. Thus, generic suffix computation can aid source-domain separability, while the training-supported transition supplies an inductive bias that is especially useful under shift.

A.3 Comparison with Internal Confidence

We evaluate Internal Confidence (Chen et al., 2026) using its official system prompt, question wrapper, final-marker token window, and full hidden-layer grid. Table 8 uses the same 3,000/1,000 Qwen3-4B examples, labels, and subject splits as the primary comparison. The published training-free aggregation is near chance on this backbone. A supervised logistic probe over its full $P(\text{Yes})$ layer–token grid reaches 0.6053/0.5910, compared with ThinkUQ at 0.7895/0.7249. The ThinkUQ prefill also improves the grid readout. The stronger hidden-state result suggests that correctness information extends beyond a single verbalized Yes/No direction.

A.4 Comparison with No Answer Needed

We reproduce the published "No Answer Needed" estimator (Moreno Cencerrado et al., 2026): at the final question token, it subtracts the midpoint of the correct/incorrect source centroids and projects the residual stream onto their unit-normalized difference. Source-validation AUROC selects the layer before target evaluation. On the full Qwen3-4B TriviaQA question-source shift (1,800 source and 600 target examples; target answer accuracy .2450), the official Query-end score reaches .7624 target AUROC. Applying the identical estimator and label budget at the ThinkUQ observation

Table 7: Primary correctness UQ under subject shift and zero-tuning transfer (Qwen3-4B; 3,000/1,000 source/target examples; answer accuracy .7207/.4320). MMLU and MMLU-Pro columns report AUROC. In the filler row, the target contrast is ThinkUQ minus filler; in the ThinkUQ row, it is ThinkUQ minus Query-end. Ex/Gp are paired hierarchical 95% example/subject CIs for the stated contrast.

Condition	MMLU	MMLU-Pro	Target Δ	Ex CI	Gp CI
Query-end state	.6900	.6524	–	–	–
Length-matched filler	.7264	.6588	$+.0661$	$[+.0371, +.0963]$	$[+.0314, +.1010]$
ThinkUQ post-boundary state	**.7895**	**.7249**	$+.0725$	$[+.0441, +.0999]$	$[+.0511, +.0929]$

Table 8: Internal Confidence on the primary Qwen3-4B MMLU subject-OOD and MMLU-Pro transfer protocol. Grid LR fits logistic regression to all layer–token $P(\text{Yes})$ values using the same source correctness labels and classifier family as THINKUQ.

Method	MMLU AUROC	MMLU-Pro AUROC
Internal Confidence (official aggregation)	.5246	.5165
Internal Confidence (naive mean)	.6002	.5245
Internal Confidence (full grid LR)	.6053	.5910
Internal Confidence + THINKUQ prefill (grid LR)	.6121	.5975
THINKUQ hidden-state probe	**.7895**	**.7249**

Table 9: Zero-tuning target results across twenty model–task points. Ex/Gp are paired 95% CIs under split+example and split+benchmark-group bootstrap. Bold deltas are nonzero under both resampling units.

Params.	Transfer task	Acc.	Query-end	THINKUQ	Gain	Ex 95% CI	Gp 95% CI
1.7B	TriviaQA	.148	.7490	.7629	$+.0139$	$[-.0184, +.0464]$	$[-.0261, +.0432]$
4B	TriviaQA	.241	.7800	.8330	$+.0530$	$[+.0271, +.0785]$	$[+.0249, +.0764]$
8B	TriviaQA	.391	.7897	.8030	$+.0133$	$[-.0028, +.0299]$	$[-.0049, +.0249]$
14B	TriviaQA	.427	.7803	.8162	$+.0359$	$[+.0190, +.0526]$	$[+.0170, +.0590]$
1.7B	PopQA	.177	.7674	.7517	$-.0156$	$[-.0691, +.0302]$	$[-.1258, +.0532]$
4B	PopQA	.202	.7422	.7688	$+.0266$	$[-.0554, +.1068]$	$[-.0664, +.1307]$
8B	PopQA	.240	.7851	.6793	$-.1059$	$[-.1637, -.0541]$	$[-.1924, -.0132]$
14B	PopQA	.277	.7245	.7613	$+.0368$	$[-.0983, +.1507]$	$[-.0989, +.1699]$
1.7B	GSM8K→MATH	.660	.5210	.5940	$+.0730$	$[-.0000, +.1463]$	$[-.0101, +.1434]$
4B	GSM8K→MATH	.742	.5758	.5980	$+.0223$	$[-.0408, +.0858]$	$[-.0434, +.0920]$
8B	GSM8K→MATH	.600	.5322	.5942	$+.0620$	$[+.0111, +.1116]$	$[-.0093, +.1368]$
14B	GSM8K→MATH	.478	.6436	.5822	$-.0614$	$[-.0946, -.0278]$	$[-.0997, -.0037]$
1.7B	GSM8K train→test	.796	.6481	.6518	$+.0037$	$[-.0347, +.0397]$	$[-.0286, +.0329]$
4B	GSM8K train→test	.890	.5655	.5711	$+.0057$	$[-.0324, +.0449]$	$[-.0108, +.0211]$
8B	GSM8K train→test	.313	.7091	.7680	$+.0589$	$[+.0349, +.0849]$	$[+.0441, +.0778]$
14B	GSM8K train→test	.346	.8162	.8280	$+.0118$	$[-.0013, +.0249]$	$[+.0082, +.0162]$
1.7B	MBPP→HumanEval	.573	.5586	.5084	$-.0501$	$[-.1308, +.0283]$	$[-.1295, +.0290]$
4B	MBPP→HumanEval	.695	.4619	.5187	$+.0568$	$[-.0150, +.1339]$	$[-.0160, +.1325]$
8B	MBPP→HumanEval	.817	.5914	.6067	$+.0152$	$[-.0647, +.1002]$	$[-.0697, +.1047]$
14B	MBPP→HumanEval	.872	.5368	.5705	$+.0338$	$[-.0647, +.1371]$	$[-.0668, +.1320]$

point reaches .7806, a point gain of $+.0182$. This gain remains unresolved under paired hierarchical percentile bootstrap: split+correctness-stratified-example CI $[-.0136, +.0506]$ ($p = .2680$) and split+question-source-group CI $[-.0103, +.0441]$ ($p = .1848$). The centroid-direction estimator is a strong Query-end baseline on this shift.

A.5 APPLICABILITY ACROSS SCALES AND TASK FAMILIES

Table 9 evaluates Qwen3-1.7B/4B/8B/14B across factual recall, arithmetic reasoning, and code generation. The tasks probe different sources of answerability: stored factual knowledge, latent multi-step solutions, and executable program structure. The cross-scale map links the boundary's readout gain to these different computational demands.

TriviaQA provides the clearest cross-scale regime: every scale has a positive point estimate, and the 4B and 14B gains resolve under both resampling units. GSM8K shows that the state can also help free-form arithmetic reasoning, most clearly at 8B. HumanEval remains unresolved across scales, consistent with correctness that depends on the complete generated program. PopQA and MATH show task-dependent responses that the class-conditional analysis below examines in more detail.

A.6 CROSS-SCALE TRIVIAQA FAMILY SYNTHESIS

We define an equal-weight family statistic over Qwen3-1.7B, 4B, 8B, and 14B on the aligned 600-example subset nested within the expanded panel. It averages the four model-specific AUROC deltas. The paired hierarchical percentile bootstrap resamples the ten source splits synchronously.

Table 10: Correctness-conditioned probability shifts at representative positive and boundary cases. AUROC pairs are Query-end/THINKUQ; deltas are THINKUQ minus Query-end. Ex/Gp are paired hierarchical split+example and split+semantic-group 95% CIs.

Point	Acc.	Query-end/THINKUQ AUROC	ΔSep.	Ex CI	Gp CI	ΔConf.-wrong	Ex/Gp CIs
14B TriviaQA	.448	.7664/.7993	$+.0525$	$[+.0310, +.0733]$	$[+.0215, +.0754]$	$-.0242$	$[-.0565, +.0085]/[-.0593, +.0186]$
8B PopQA	.240	.7851/.6793	$-.1796$	$[-.2507, -.1188]$	$[-.2571, -.0235]$	$-.0329$	$[-.0925, +.0390]/[-.2909, +.1049]$

The example analysis also resamples target questions synchronously within the joint four-model correctness patterns; the group analysis instead resamples the seven question sources synchronously. The fixed seed is 2026081301. This statistic measures the mean cross-scale TriviaQA effect.

At the primary shared source-selected layer, the four model deltas are $+.0570$, $+.0009$, $+.0196$, and $+.0354$, giving a family effect of $+.0282$. Its example CI is $[+.0113, +.0455]$ and its group CI is $[+.0176, +.0368]$; neither bootstrap has a nonpositive draw ($p < .0004$ at 5,000-draw resolution). The conclusion survives both layer sensitivities. At condition-specific source-selected layers, the family effect is $+.0357$, with example/group CIs $[+.0181, +.0536]$ ($p = .0008$) and $[+.0243, +.0460]$ ($p < .0004$). At the globally fixed 75%-depth layer, it is $+.0240$, with CIs $[+.0054, +.0425]$ ($p = .0116$) and $[+.0096, +.0365]$ ($p = .0056$). Thus all four checkpoints contribute to a positive family-level estimate under both resampling units and all three layer definitions.

A.7 CORRECTNESS-CONDITIONED SHIFTS IDENTIFY THE POSITIVE REGIME

We analyze the ten frozen source-fit probability vectors by answer correctness and the target's predefined semantic groups, using fixed subgroup definitions and thresholds. Table 10 reports the change in probability separation, defined as the THINKUQ-minus-Query-end shift for correct answers minus that for incorrect answers, and the change in the fixed-threshold confident-wrong rate among incorrect answers. Positive separation is favorable; positive confident-wrong change is adverse.

TriviaQA's favorable separation is symmetric: the boundary raises mean correct-answer probability by $+.0270$ and lowers incorrect-answer probability by $-.0255$, with both shifts significant under both units. On PopQA it instead reduces correct-answer probability by $-.1732$ while leaving the incorrect shift near zero, producing the adverse separation result. The contrast links the positive ranking effect to an increased gap between correct and incorrect examples, rather than a common shift in their predicted confidence.

A few-label experiment fits the probe with 4, 8, or 16 balanced source labels per split. The readout uses the globally fixed 75%-depth layer (raw layer 29), $C = 1$, the ten archived source splits, and seed 20260826. On Qwen3-14B TriviaQA ($n = 600$, accuracy .448), THINKUQ exceeds Query-end at all three budgets: deltas are $+.0401$, $+.0876$, and $+.0758$, each positive under both paired split+example and split+group bootstrap. The four-label interval is already $[+.0075, +.0748]/[+.0010, +.0718]$. The positive boundary gain is therefore accessible with a small probe-fitting budget once the configuration is fixed. Appendix F specifies the source pool and separates this fit budget from source-side configuration assessment.

A label-free diagonal-Gaussian typicality score remains near chance on the same Qwen3-14B TriviaQA regime: Query-end/THINKUQ AUROC is $.5009/.5041$ ($\Delta = +.0032$), with paired fold+example and fold+group CIs $[-.0610, +.0661]/[-.0613, +.0578]$. This unresolved contrast, alongside the few-label gains, highlights the value of correctness supervision for the readout.

A.8 PROBABILITY CALIBRATION AND SELECTIVE PREDICTION

We evaluate probability calibration and selective prediction with frozen source-selected probes and no target recalibration, reporting NLL and risk-at-50%-coverage contrasts in Table 11. On Qwen3-14B TriviaQA, THINKUQ improves NLL under example resampling and leaves selective risk unresolved; on Qwen3-8B PopQA, the adverse ranking regime also raises selective risk. These metrics distinguish the readout's ranking quality from the probability estimates and selective decisions obtained from its scores.

Table 11: Frozen source-to-target calibration and selective risk at a positive regime and a task boundary. AUROC pairs are Query-end/THINKUQ; loss deltas are THINKUQ minus Query-end, so negative is better. Ex/Gp are paired hierarchical split+example and split+semantic-group 95% CIs.

Point	Acc.	AUROC pair	ΔNLL	Ex CI	Gp CI	ΔRisk@50	Ex CI	Gp CI
14B TriviaQA	.448	.7664/.7993	$-.0344$	$[-.0607, -.0083]$	$[-.0660, +.0044]$	$-.0237$	$[-.0490, +.0017]$	$[-.0593, +.0121]$
8B PopQA	.240	.7851/.6793	$-.1683$	$[-.3367, -.0296]$	$[-.6597, +.0805]$	$+.0600$	$[+.0267, +.0980]$	$[+.0007, +.1233]$

Table 12: Boundary and readout-depth ablations. "Post-boundary state 1" is the THINKUQ observation point: the first state after the complete empty reasoning boundary. States 2/3 append one/two additional fixed newline readouts.

Observation point	MMLU	MMLU-Pro	Structural condition	MMLU	MMLU-Pro
Exact `<think>`	.6484	.6545	No tags, same layout	.7949	.7130
Exact `</think>`	.6922	.6655	Open tag only	.7910	.7210
Post-boundary state 1	**.7886**	**.7254**	Close tag only	.7786	.7167
Post-boundary state 2	.7835	.7182	Swapped tags	.7836	.7153
Post-boundary state 3	.7784	.7177	Correct full boundary	.7886	.7254

Target ECE is $.0601/.0542$ for Query-end/THINKUQ on TriviaQA and $.2031/.1974$ on PopQA; both ECE deltas are unresolved under the required agreement of example and group resampling. Ranking, probability calibration, and selective risk describe distinct uses of the same correctness scores.

B WHAT DOES THE EMPTY BOUNDARY CHANGE?

B.1 THE SIGNAL IS TOKEN-LOCALIZED

Table 12 traces the signal through the fixed suffix. The exact opening and closing tag states are weak. Performance peaks at the first common token *after* the complete boundary and degrades when one or two additional fixed readout tokens are appended. The localization is important: treating the diagnostic text as one segment and averaging it would obscure the best state. It also mirrors answer-level findings that correctness signals concentrate at semantically privileged token positions (Orgad et al., 2025), but here the privileged position exists before any answer is generated.

B.2 THE NATIVE BOUNDARY BEATS RANDOM ORDINARY SUFFIXES

We evaluate a representative Qwen3-14B TriviaQA control at raw layer 29 with fixed $C = .001$, 400 source examples, and 200 target examples. Twenty four-token ordinary suffixes are sampled once and shared across examples; a second intervention moves the exact four boundary-token IDs inside the user message. Target accuracy is .4900. Query-end, THINKUQ, and alternative-position AUROC are .7486, .7811, and .7722, while the ordinary-suffix mean/median/range is .6973/.7017/[.6382,.7379]. None of the 20 ordinary draws reaches the THINKUQ point estimate.

The primary contrast, THINKUQ minus the random-suffix mean, is $+.0837$: its paired split+example CI is $[+.0433, +.1274]$ ($p = .0010$) and split+group CI is $[+.0354, +.1243]$ ($p = .0020$), with suffix draws resampled inside both bootstraps. THINKUQ minus Query-end is $+.0324$, with an example interval crossing zero ($[-.0071, +.0758]$, $p = .1119$) and a positive group interval ($[+.0007, +.0561]$, $p = .0460$), leaving this contrast unresolved under the both-unit rule. Moving the same boundary IDs gives $+.0235$ over Query-end, with example/group CIs $[-.0188, +.0682]$/$[-.0113, +.0458]$, both unresolved. The native boundary's advantage over ordinary suffixes shows that equal token count alone does not account for the stronger readout.

B.3 OFFICIAL TEMPLATE MODES RECOVER THE INTERVENTION EXACTLY

We render five modes through the checkpoint's official Qwen3 chat template on a fixed, subject-covered Qwen3-4B MMLU→MMLU-Pro sample (400 source/200 target; target accuracy .4200). Token-level checks show that `enable_thinking=true` with an assistant-generation prompt exactly recovers the Query-end condition, whereas `enable_thinking=false` exactly appends

the four empty-boundary tokens and recovers the THINKUQ condition. Every mode decodes zero tokens. Ten source-subject folds select layer and regularization by source-validation NLL before target transfer.

Under canonical probe transfer, AUROC for assistant-generation/official-no-thinking/user-turn-end/fixed-`Answer:`-prefill/closed-empty-assistant is $.5754/.6683/.5501/.6997/.6620$. Official no-thinking minus assistant-generation is $+.0928$, with paired hierarchical split+correctness-stratified-example CI $[+.0257, +.1648]$ ($p = .0090$) and split+subject CI $[+.0168, +.1739]$ ($p = .0160$), positive under both units. The other frozen transfer contrasts are unresolved: user-turn-end minus generation is $-.0253$ (CIs $[-.0710, +.0212]$/$[-.0845, +.0317]$), fixed answer prefix minus no-thinking is $+.0315$ ($[-.0069, +.0702]$/$[-.0106, +.0688]$), and closed-empty assistant minus no-thinking is $-.0063$ ($[-.0479, +.0354]$/$[-.0525, +.0409]$).

When each mode instead receives its own source-only selected probe, the same five AUROCs are $.5754/.6683/.6685/.6803/.6365$. The canonical no-thinking gain remains $+.0928$ (example/subject CIs $[+.0271, +.1638]$/$[+.0211, +.1674]$), and user-turn-end becomes $+.0931$ over generation ($[+.0333, +.1589]$/$[+.0251, +.1803]$); answer-prefix and closed-assistant differences from no-thinking remain unresolved. The reversal for user-turn-end across probe regimes shows that a template boundary can change both information accessibility and which source-trained readout transfers. The exact token equality identifies THINKUQ with the model's native no-thinking transition.

B.4 THE NATIVE BOUNDARY OUTPERFORMS A SOURCE-LEARNED SOFT PREFIX

We replace the four native boundary tokens by exactly four trainable continuous embeddings, initialized from those same four input-embedding rows. On the frozen Qwen3-4B MMLU→MMLU-Pro protocol (400 source/200 target; target accuracy $.4200$), the backbone is fixed and the prefix and a temporary checkpoint-selection head are optimized. A single deterministic source-subject split supplies 234 training and 78 validation examples; source-validation NLL selects step 70 of 120 at fixed raw layer 26. We then replace the temporary head with separate capacity-matched source-only probes for Query-end, THINKUQ, and the learned prefix.

Target AUROC is $.5452/.7108/.5446$ for Query-end/THINKUQ/learned soft prefix. THINKUQ minus Query-end is $+.1656$, with paired correctness-stratified-example percentile CI $[+.0898, +.2414]$ ($p = .0010$) and paired subject-cluster percentile CI $[+.0745, +.2408]$ ($p = .0020$), positive under both units. The learned prefix minus THINKUQ is $-.1662$, with example/subject CIs $[-.2422, -.0969]$/$[-.2338, -.1063]$ (both $p = .0010$), negative under both units. The learned prefix minus Query-end is $-.0005$, with CIs $[-.0625, +.0572]$/$[-.0689, +.0532]$ ($p = .9825/.9455$), unresolved under both units. The native discrete boundary therefore provides a stronger readout than the source-trained continuous prefix at matched token count.

B.5 THE LAYER TRAJECTORY IS TASK DEPENDENT

Figure 3 summarizes the complete nine-slot target curves for three representative tasks. The curves are descriptive ten-fold means; the vertical marker identifies the globally fixed 75%-depth slot used for paired inference. On Qwen3-14B TriviaQA (answer accuracy .4483), fixed-layer Query-end/THINKUQ AUROC is $.7572/.7948$, a $+.0376$ delta with paired split+example CI $[+.0137, +.0632]$ ($p = .0020$) and split+group CI $[+.0015, +.0738]$ ($p = .0412$), positive under both units. On Qwen3-8B MATH (accuracy .6000), the corresponding $.5552/.5865$ delta is $+.0313$, with example/group CIs $[-.0228, +.0846]$/$[-.0455, +.1165]$, both unresolved. On Qwen3-8B PopQA (accuracy .2400), it is $.8280/.7075$, a $-.1205$ point estimate; the example CI $[-.1879, -.0587]$ is negative, but the group CI $[-.2224, +.0028]$ crosses zero, leaving the both-unit result unresolved. These curves reveal task-dependent depth profiles. The geometry analysis below characterizes the corresponding state changes, and Appendix D examines token-by-layer responses, attention routing, and residual-state interventions.

B.6 REORGANIZATION MAGNITUDE DOES NOT EXPLAIN THE UQ EFFECT

Figure 4 shows centered linear CKA between Query-end and THINKUQ states and between Query-end and the equal-token-count filler across all nine layers of the same three panels. At the globally fixed 75%-depth layer, Query-end–THINKUQ/Query-end–filler CKA is $.0479/.0729$ on Qwen3-14B

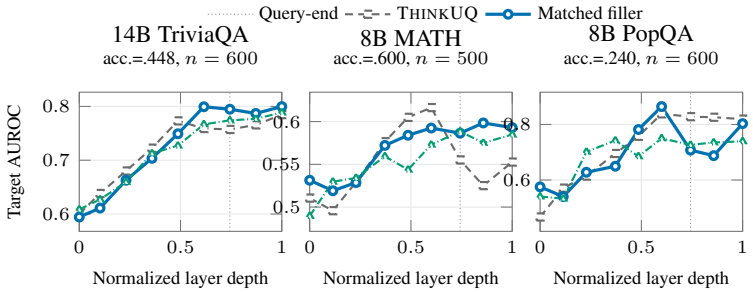

Figure 3: Representative target AUROC trajectories across nine residual-stream depths. Each curve is the mean of ten frozen source folds; Query-end, THINKUQ, and filler use separately source-fit probes at each displayed layer. The vertical line marks the globally fixed 75%-depth layer used for the paired split+example and split+group comparisons reported in the text.

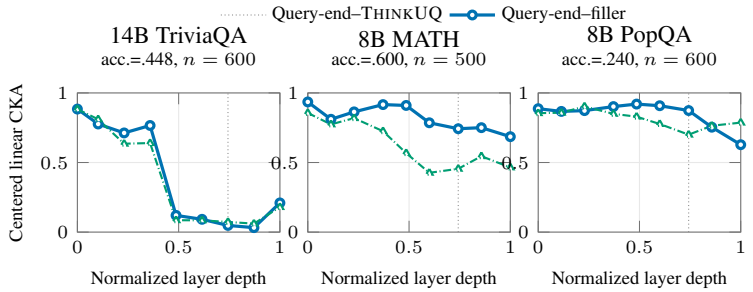

Figure 4: Representative residual-stream geometry using all target examples (TriviaQA $n = 600$, MATH $n = 500$, PopQA $n = 600$). Curves are descriptive centered linear CKA of Query-end with THINKUQ or equal-length filler states at nine depths. The vertical line marks the globally fixed 75%-depth layer used for the paired normalized-L2 comparisons reported in the text.

TriviaQA, .7428/.4549 on Qwen3-8B MATH, and .8736/.7002 on Qwen3-8B PopQA. Thus only TriviaQA makes the THINKUQ state descriptively less Query-end-like than filler in the CKA sense; the other two reverse that ordering.

We compare per-example normalized L2 shifts from Query-end at the fixed layer. The THINKUQ-minus-filler shift is $-.0111$ on TriviaQA, with correctness-stratified paired example CI $[-.0114, -.0107]$ and paired group CI $[-.0114, -.0106]$; it is $-.0273$ on MATH, with CIs $[-.0293, -.0252]$/$[-.0315, -.0213]$; and $+.0547$ on PopQA, with CIs $[+.0536, +.0557]$/$[+.0505, +.0573]$. All three exclude zero under both units (zero empirical-tail draws, $p < .001$). Their directions differ from the UQ effects: TriviaQA combines a positive ranking gain with a smaller displacement, MATH combines unresolved UQ with a smaller displacement, and the adverse PopQA point has the larger displacement. This pattern favors an account based on the direction and organization of the state change, rather than displacement magnitude alone.

B.7 THE FIXED READOUT WIDENS ITS LINEAR MARGIN SELECTIVELY

We analyze the ten source-fit probability vectors at a fixed layer with exact target-ID alignment. For each source split and condition, we convert correctness probabilities to logits and measure correct-minus-incorrect separation in pooled within-class standard-deviation units. Table 13 reports the THINKUQ-minus-Query-end change at the same globally fixed 75%-depth layer. TriviaQA's standardized separation increases from .9958 to 1.1813 ($\Delta = +.1855$), with paired split+example CI $[+.0668, +.3017]$ ($p = .0040$) and split+group CI $[+.0341, +.3161]$ ($p = .0200$), positive under both units. MATH changes from .2241 to .2678 ($\Delta = +.0437$), with CIs $[-.1575, +.2272]$/$[-.1908, +.3377]$, both unresolved. PopQA falls from 1.3832 to .8928 ($\Delta = -.4904$): the example CI $[-.7987, -.2095]$ is negative, but the group CI $[-.9659, +.0890]$

Table 13: Frozen fixed-layer linear-readout separation at 75% model depth. Margin is correct-minus-incorrect logit separation divided by pooled within-class SD, averaged over ten source splits. Intervals are paired 95% split+example and split+group percentile CIs for THINKUQ minus Query-end; significance requires both units.

Panel	Query-end	THINKUQ	Δ	Example CI	Group CI
14B TriviaQA	.9958	1.1813	$+.1855$	$[+.0668, +.3017]$	$[+.0341, +.3161]$
8B MATH	.2241	.2678	$+.0437$	$[-.1575, +.2272]$	$[-.1908, +.3377]$
8B PopQA	1.3832	.8928	$-.4904$	$[-.7987, -.2095]$	$[-.9659, +.0890]$

crosses zero, leaving the both-unit result unresolved. For PopQA, 295 of the 6,000 Query-end fold–example probabilities reach the numerical clipping boundary. Clipping affects its logit-scale magnitude; the rank-based AUROC is computed from the original probabilities.

The aligned fixed-layer AUROC points are the values reported above: TriviaQA $.7572/.7948$ ($\Delta = +.0376$, positive under both units), MATH $.5552/.5865$ ($\Delta = +.0313$, unresolved), and PopQA $.8280/.7075$ ($\Delta = -.1205$, unresolved under both units). The label-efficiency curves reach 95% of the full excess at 256 and 64 labels for TriviaQA and MATH. PopQA reaches its lower full-readout threshold at 32 labels along a non-monotone curve; class counts make the 512-label point infeasible. These results describe how fixed-layer separation and target ranking respond to the probe-fitting budget.

Prequential coding measures sequential predictability from source data. A frozen, correctness-stratified online order uses cumulative endpoints $10/20/40/70/100\%$: the first block receives one uniform bit per label, while each later block is encoded by a fixed-$C = 1$ linear probe trained only on the preceding blocks. Every panel uses all 1,800 source examples, the same blocks for Query-end/THINKUQ/filler. THINKUQ-minus-Query-end bits/example is $-.4401$ on TriviaQA (Query-end/THINKUQ $2.0823/1.6423$), with paired source-example CI $[-.5957, -.2787]$ ($p = .0010$) and seven-subject CI $[-.7012, -.2045]$ ($p = .0010$): THINKUQ is shorter under both units. MATH gives $-.0666$ ($2.0572/1.9906$), with example CI $[-.2562, +.1036]$ and five-online- block cluster CI $[-.2277, +.1174]$, both unresolved. Its GSM8K source forms a single semantic group, so cluster resampling uses the five online blocks. PopQA gives $+.1798$ ($1.0919/1.2717$), with example CI $[+.0565, +.3040]$ ($p = .0060$) and seven-subject CI $[+.0650, +.3689]$ ($p = .0070$): THINKUQ is longer under both units. Absolute codelengths exceed the one-bit uniform code; the paired contrast measures relative predictive efficiency along the same online order. Its signs agree with the positive, unresolved, and adverse target results, connecting source coding efficiency to transferred readability.

B.8 THE GAIN SURVIVES A MATCHED PROMPT PARAPHRASE

We evaluate a representative Qwen3-4B MMLU→MMLU-Pro surface test with 400 source and 200 target examples; the answer policy is unchanged and target accuracy is .4450. Under matched-surface fitting, the canonical Query-end/THINKUQ AUROC is .6061/.6867 ($\Delta = +.0807$), with paired fold+example CI $[+.0196, +.1414]$ and fold+group CI $[+.0012, +.1615]$. A deterministic instruction and field-label paraphrase gives .5772/.6737 ($\Delta = +.0965$), with CIs $[+.0249, +.1735]$/$[+.0085, +.1920]$; both gains are positive under both units. Deterministically permuting all demonstration and test options while preserving answer text gives .6050/.6441 ($\Delta = +.0392$), unresolved under both units. The four perturbation-minus-canonical difference-in-differences are also unresolved under both units: matched paraphrase/permutation are $+.0158/-.0415$, and zero-adaptation canonical-probe transfer gives $-.0025/-.0156$. Table 19 reports all contrasts. Matched paraphrasing preserves the positive boundary gain; the frozen-probe transfer contrasts remain unresolved under the both-unit rule.

B.9 A SINGLE-STATE READOUT CAPTURES THE BOUNDARY BENEFIT

We evaluate three-anchor fusion, query/tag interactions, two- and three-layer windows, a cross-fitted layer stacker, and a PCA-bottleneck MLP. None simultaneously and significantly improves both source and target over the single-state probe. A three-anchor ensemble reaches 0.7304 on MMLU-Pro but falls to 0.7723 on MMLU; a three-layer state window reaches 0.7897/0.7260, statistically

indistinguishable from the primary readout; and the PCA–MLP regresses to 0.7565/0.6867. Within these boundary-feature designs, the single-state linear readout captures the benefit without additional fusion or nonlinear probe capacity.

B.10 Counterfactual response and layer geometry

Raw flattened $(\text{think} - \text{filler})$ differences do not transfer reliably. Compact layerwise summaries of the displacement do: CCPS reaches 0.7208 AUROC on MMLU-Pro, and the strongest component describes MLP-write geometry rather than logit-lens shift. Combining CCPS with the raw ThinkUQ state improves ranking modestly in some protocols but does not beat the strong single-state result across both domains. The MLP-write summaries connect the prefill effect to changes in internal computation.

C A Readout-Geometry Account of ThinkUQ

ThinkUQ uses a native computational transition to expose correctness information to a lightweight readout. We analyze this operation through three complementary quantities: information in the input, class separation in the representation, and the predictive loss of a fitted probe. This connects the observation-state intervention to the ranking, margin, and codelength results.

A fixed boundary changes the representation. Let Q be the complete query, $Y \in \{0, 1\}$ its eventual correctness label under a fixed answer policy, and B the deterministic empty-boundary sequence. For a frozen model, let $H_q = f_q(Q)$ and $H_b = f_b(Q, B)$ denote the Query-end and post-boundary states at a fixed layer. Because B is constant and both states are deterministic functions of Q,

$$I(Y; Q, B) = I(Y; Q), \qquad I(Y; H_b) \le I(Y; Q). \tag{5}$$

The boundary therefore operates on existing query information. Its hidden state is computed by nonlinear Transformer layers with attention to the full prompt, giving it access to representations distributed across earlier positions. In particular, H_b is a new function of Q, rather than necessarily a function of the single vector H_q.

What a linear observer can recover. For affine scores $s_{w,c}(h) = w^\top h + c$, define the population logistic readout risk

$$\mathcal{R}_{\text{lin}}(H) = \inf_{w,c} \mathbb{E}\Big[\log\Big(1 + \exp\{-\tilde{Y}\, s_{w,c}(H)\}\Big)\Big], \qquad \tilde{Y} = 2Y - 1. \tag{6}$$

Lower $\mathcal{R}_{\text{lin}}$ means better probabilistic prediction by an optimal affine-logistic observer. AUROC instead measures pairwise ranking. These objectives capture different aspects of readout quality, which we evaluate separately through NLL and AUROC.

If $H_b = AH_q + a$ for a fixed matrix A, every affine score of H_b is also an affine score of H_q:

$$w^\top H_b + c = (A^\top w)^\top H_q + (w^\top a + c). \tag{7}$$

Consequently, $\mathcal{R}_{\text{lin}}(H_q) \le \mathcal{R}_{\text{lin}}(H_b)$, with equality when A is invertible. The same inclusion of score families applies to optimal affine-score AUROC. A fixed linear projection can improve finite-sample estimation through regularization, but cannot enlarge this population score family. ThinkUQ provides a richer operation: nonlinear computation at a new position can reorganize or collect prompt information into an affine-readable state. Matching probe families and label budgets isolates the empirical contribution of that observation state.

Class separation determines ranking in a Gaussian score model. Suppose $H \mid Y = k \sim \mathcal{N}(\mu_k, \Sigma)$, with shared positive-definite covariance Σ, and write $\Delta\mu = \mu_1 - \mu_0$. For independent correct and incorrect examples, their score difference along a nonzero direction w is Gaussian with mean $w^\top \Delta\mu$ and variance $2w^\top \Sigma w$. Hence

$$\text{AUROC}(w) = \Phi\left(\frac{w^\top \Delta\mu}{\sqrt{2w^\top \Sigma w}}\right), \tag{8}$$

where Φ is the standard-normal CDF. The Cauchy–Schwarz inequality gives the maximizing direction $w \propto \Sigma^{-1}\Delta\mu$ when $\Delta\mu \neq 0$, and

$$d^2 = \Delta\mu^\top \Sigma^{-1}\Delta\mu, \qquad \mathrm{AUROC}^\star_{\mathrm{lin}} = \Phi\left(\frac{d}{\sqrt{2}}\right). \tag{9}$$

This relation explains why the direction of a representation change matters more than its magnitude. Increasing correctness separation relative to within-class variation improves linear ranking in this model. The fixed-layer standardized logit margin measures the corresponding separation of the fitted score, connecting the positive TriviaQA margin to its higher AUROC.

The source direction must retain target separation. For each observation state, let w_s be its source-fitted direction and define

$$\gamma_t(w_s) = \frac{w_s^\top(\mu_{t,1} - \mu_{t,0})}{\sqrt{w_s^\top(\Sigma_{t,1} + \Sigma_{t,0})w_s}}. \tag{10}$$

Here $\mu_{t,k}$ and $\Sigma_{t,k}$ are target class means and covariances, and the score variance in the denominator is positive. Gaussian target score distributions give $\mathrm{AUROC}_t = \Phi(\gamma_t)$ even with unequal class variances. Thus, under this score model, a transferred boundary advantage is exactly a larger γ_t for its own frozen source direction. Shared- and fixed-layer comparisons evaluate whether the native transition improves this readability within the same representation depth.

Prequential coding measures sequential predictability. For the source order used in the coding experiment, let $\widehat{p}_{<i}$ denote the predictor trained on preceding blocks, with uniform probability $1/2$ for the initial block. Its codelength is

$$L = -\sum_i \log_2 \widehat{p}_{<i}(Y_i \mid H_i). \tag{11}$$

A shorter code means that the sequence of fitted probes assigns higher geometric-mean probability to the realized labels. It therefore measures learning from finite source prefixes, complementing the target ranking and margin diagnostics. On TriviaQA, the lower source codelength, wider fixed-layer margin, and higher transferred AUROC provide convergent evidence of improved correctness readability. The Tagged-CoT trajectory then tracks how the relative ranking advantage develops within one shared layer. Together, these analyses connect a native state transition to the efficiency with which a simple observer extracts correctness information.

D Training-Time Trajectories of the Empty-Boundary UQ Signal

We study training-time acquisition from the released non-thinking Qwen3-4B-Instruct-2507 checkpoint, holding the readout layer fixed throughout Tagged-CoT continuation.

D.1 A fixed-layer advantage emerges from the released non-thinking checkpoint

Starting from Qwen3-4B-Instruct-2507, we train a fresh rank-16 LoRA on 9,984 examples from CoT Collection (Kim et al., 2023), serialized with the official Qwen reasoning-content tags. We preserve the initial checkpoint and evaluate every 30 minutes through 600 minutes. Each checkpoint uses the same 1,200-example MMLU source manifest, 600-example MMLU-Pro target manifest, source-only probe fit, and common raw layer 30 for Query-end and ThinkUQ. The matched layer and fixed evaluation pools track the relative readout advantage at the same observation depth throughout training.

Training uses four data-parallel workers, one sequence per worker and eight accumulation steps, for a global batch of 32. LoRA has $\alpha = 32$ and dropout .05 and updates the attention q/k/v/o and MLP gate/up/down projections. AdamW uses betas $(.9, .95)$, zero weight decay, and gradient-norm clipping at 1.0. The learning rate warms up over the first 3% of wall-clock training time to 5×10^{-5}, then follows cosine decay to 5×10^{-6}. Maximum sequence length is 1,024 and the fixed seed is 20260826. The horizontal axis in Figure 2 records wall-clock minutes for this four-worker configuration.

Table 14: Token-by-layer response on Qwen3-4B MMLU-Pro ($n = 200$, 14 subjects; seed 20260909). Values are native-boundary-minus-filler contrasts at fixed raw layer 26 and the fourth suffix token. Ex/Gp are paired 95% correctness-stratified example/subject-cluster percentile CIs. The table reports inference at the predefined primary location; the complete 36-layer, five-position arrays describe the response across depth and token position.

Metric	Δ	Ex 95% CI	Gp 95% CI	Both-unit call
Normalized residual displacement	$+.0711$	$[+.0695, +.0728]$	$[+.0678, +.0739]$	Positive
Attention-write RMS	$+.0442$	$[+.0378, +.0506]$	$[+.0273, +.0617]$	Positive
MLP-write RMS	$-.0162$	$[-.0215, -.0104]$	$[-.0267, -.0038]$	Negative
Attention mass: recent 16 query tokens	$-.0543$	$[-.0566, -.0521]$	$[-.0598, -.0491]$	Negative
Attention mass: visible suffix	$-.0323$	$[-.0339, -.0306]$	$[-.0356, -.0288]$	Negative
Gold-choice logit margin	$-.5947$	$[-1.0793, -.0993]$	$[-1.1867, +.0248]$	Unresolved

Figure 2 shows the complete 21-checkpoint target curve. The initial $+.0080$ difference is unresolved (95% CI $[-.0226, +.0420]$), whereas the 600-minute difference is $+.1160$ (CI $[+.0557, +.1774]$, two-sided $p = .0010$). Every point estimate is positive and 16 of 21 intervals exclude zero. Along this training path, training time and ΔAUROC have Spearman $\rho = .660$ (nominal $p = .0011$). The coefficient summarizes the shared-evaluation trajectory; pointwise intervals use paired correctness-stratified example resampling. The increasing relative advantage is therefore visible within one shared representation depth throughout continuation.

D.2 Token-by-layer response at a fixed location

We examine the computation at every layer and at the query anchor plus all four suffix positions on a fixed sample of 200 Qwen3-4B MMLU-Pro items spanning all 14 subjects. Table 14 reports six prespecified contrasts between the empty boundary and a four-token period filler at raw layer 26 and the final suffix token. Relative to filler, the native boundary produces greater normalized residual displacement and a larger attention-sublayer write, but a smaller MLP write and less attention mass on both the last 16 query tokens and the visible suffix. These five differences survive both paired correctness-stratified example bootstrap and subject-cluster bootstrap. The gold-choice logit-margin contrast is unresolved under the both-unit rule. The inputs share the exact query prefix, decode no tokens, and reconstruct attention with error below 10^{-6}. Thus the boundary induces a robust local computational response without directly forcing the output logits toward the gold choice.

D.3 Correctness-conditioned attention routing

We analyze attention routing on a separate fixed set of 100 Qwen3-4B MMLU examples (69 correct and 31 incorrect), comparing the native post-boundary readout with an equal-length period filler. For each attention statistic, the correctness interaction is the mean post-boundary-minus-filler change on correct examples minus the corresponding change on incorrect examples. To examine whether the pattern extends beyond a single layer, we split the examples deterministically into discovery (49 examples) and validation (51 examples), select five heads for each of five nonredundant statistic families using only discovery data, and evaluate the resulting 25 head–statistic pairs on validation data. Twenty-one pairs retain the discovery direction, and 10 have validation percentile intervals excluding zero. Seven of these ten lie in layers 20–31, covering layers 20, 24, 26, 28, and 31; the remaining three form a smaller early-layer cluster at layers 8–9. Thus the recurring signal is concentrated in a middle-to-late processing band rather than being unique to layer 26.

Table 15 reports the head-averaged layer-26 slice, where the routing pattern is most coherent across complementary statistics. The post-boundary state increases total prompt mass while reducing mass on the immediately preceding 32–128 prompt tokens. Its attention reaches farther back on average and becomes more concentrated. The correctness interactions sharpen the same account: correct examples show larger increases in concentration, maximum weight, mean distance, and global prompt mass, together with larger reductions in local-tail mass. These correctness-conditioned associations support selective aggregation of distributed prompt information. The full-residual intervention below complements this attention pattern by testing the readout's sensitivity to replacement of its internal state.

Table 15: Attention-routing changes at representative raw layer 26 on Qwen3-4B MMLU ($n = 100$). Δ is post-boundary minus equal-length filler. Interaction is the Δ difference between correct and incorrect examples. Intervals are paired 95% correctness-stratified example percentile intervals.

Attention statistic	Δ	95% CI	Correctness interaction	95% CI
Prompt mass	$+.0593$	$[+.0556, +.0628]$	$+.0186$	$[+.0100, +.0274]$
Recent 32-token mass	$-.0969$	$[-.1021, -.0917]$	$-.0281$	$[-.0388, -.0174]$
Recent 64-token mass	$-.0708$	$[-.0768, -.0650]$	$-.0213$	$[-.0331, -.0092]$
Recent 128-token mass	$-.0505$	$[-.0564, -.0445]$	$-.0217$	$[-.0346, -.0091]$
Mean distance	$+.0914$	$[+.0840, +.0985]$	$+.0428$	$[+.0289, +.0572]$
Maximum weight	$+.0161$	$[+.0067, +.0247]$	$+.0428$	$[+.0248, +.0613]$
Concentration	$+.0257$	$[+.0155, +.0357]$	$+.0446$	$[+.0239, +.0655]$

Residual-state swapping. We intervene on the Qwen3-4B MMLU→MMLU-Pro transfer setting by swapping the full residual stream at fixed raw layer 18 between exactly paired Query-end and THINKUQ passes. The sample is selected independently of correctness labels and contains 200 items from all 14 subjects (answer accuracy .4550); each of the ten folds reuses its source-fitted probes and frozen readout choices. Patching the Query-end state into the THINKUQ pass reduces mean AUROC from its identity-patch value .6803 to .5074, a $-.1729$ change with paired hierarchical split+example CI $[-.2732, -.0691]$ ($p < .001$) and split+subject CI $[-.2578, -.0962]$ ($p < .001$), negative under both units. The reverse THINKUQ-into-Query-end patch also decreases AUROC, from .6088 to .5318 ($\Delta = -.0769$): its example CI $[-.1683, +.0046]$ ($p = .0680$) crosses zero while its subject CI $[-.1720, -.0014]$ ($p = .0470$) is negative, leaving this contrast unresolved under the both-unit rule. Both directions reduce the paired gold-choice margin (-1.6851 and -1.2798; both example and subject intervals exclude zero). Identity patches reproduce the original states, choice logits, and all ten target AUROCs exactly. The two-way disruption is consistent with compatibility between the residual state and subsequent computation, rather than a freely transplantable scalar UQ signal. Together with the token and suffix controls, this intervention links the readout gain to the native computation elicited at the boundary.

E SCOPE

The evidence spans four Qwen3 scales, multiple-choice, factual, arithmetic, and code tasks, together with token, suffix, template, prefix, layer, and prompt-surface controls. The strongest positive evidence appears on MMLU transfer and across Qwen3 scales on TriviaQA; arithmetic and code results clarify when correctness is already readable before generation.

The observer uses white-box states and labeled source correctness. Probability calibration and selective risk are evaluated in Appendix A.8. The latency measurement uses Qwen3-14B on TriviaQA with a canonical four-token suffix. Token-position, ordinary-suffix, official-template, and prompt-surface controls characterize the role of the native boundary and its surrounding instruction.

F DETAILED EXPERIMENTAL PROTOCOL

F.1 BACKBONE, ANSWER POLICY, AND DATASETS

The primary frozen-model study uses the released thinking-capable Qwen3-4B checkpoint (Yang et al., 2025). The training-acquisition trajectory starts from the official non-thinking Qwen3-4B-Instruct-2507 checkpoint. MMLU correctness labels are produced by a fixed five-shot constrained policy over A/B/C/D, and MMLU-Pro uses the analogous A–J policy. Bare-letter and leading-space tokenizations are aggregated to reduce answer-format noise. The answer pass always receives the original prompt. Observer branches never receive the selected answer, correctness label, or choice distribution.

Table 1 uses 3,000 MMLU source and 1,000 MMLU-Pro target questions. Ten frozen folds partition source subjects into 60/20/20 train/validation/test groups; a subject occurs in exactly one group within a fold. Table 2 uses 1,800 source examples for each TriviaQA and GSM8K model, with 1,200 and 1,319 target examples, respectively. TriviaQA source and target question sites are disjoint; GSM8K transfers from its official train split to its official test split. The HumanEval comparison uses 427

sanitized MBPP source tasks and all 164 HumanEval targets. Within the source pools of Table 2, correctness-stratified 60/20/20 splits supply ten probe-fitting folds.

Free-form answers use deterministic generation through the official no-thinking chat template. TriviaQA allows 64 output tokens and uses normalized exact match against official aliases. GSM8K allows 2,048 tokens, requests a boxed final answer, and uses symbolic equivalence with normalized exact-match fallback. HumanEval allows 512 tokens and uses execution-defined correctness in an isolated worker. Outputs that reach the token limit before EOS are labeled incorrect. The same fixed answers and labels are shared by all observer conditions in each comparison.

The cross-scale family analysis in Appendix A.6 uses the aligned 600-target TriviaQA subset. Tables 4 and 5 use Qwen3-14B with 1,800 source and 600 target TriviaQA examples and ten question-source-disjoint fitting folds. Table 6 uses a paired Qwen3-4B TriviaQA answer subset and its archived source-fitted observer scores; the latency panel uses 64 paired Qwen3-14B queries. Table 3 uses control-specific pools and readouts detailed in Appendices B and D. Figure 2 uses a separate fixed 1,200-example MMLU and 600-example MMLU-Pro set at every checkpoint, with common raw layer 30 throughout.

F.2 Readout selection and source-side use

The source-side diagnostic is a configuration step on labeled source data: fit matched Query-end and post-boundary probes on source training examples, compare their validation ranking and correct–incorrect separation, then freeze the chosen observation point and readout for new queries. The state standardizer and probe are fitted on source data; layer and regularization selection use source-validation NLL. Target labels enter the final evaluation. This separates choosing an observation configuration from fitting the probe at a specified label budget.

Matched single-state probes in the primary and cross-scale panels use an inverse-regularization grid $C \in \{.001, .01, .1, 1\}$ and nine evenly spaced residual-stream depths. Condition-specific selection chooses a layer and C for each observer using source validation. Shared-layer analyses compare observers at one common source-selected layer, while globally fixed-layer analyses use the 75%-depth slot. In Table 4, target AUROC and probability separation use condition-specific source-selected readouts; margin uses the fixed-depth readout, and prequential coding fits $C = 1$ probes on successive source prefixes. These diagnostics measure complementary properties of the same paired states.

Table 5 fixes raw layer 29 (75% depth), $C = 1$, ten source splits, and seed 20260826 before fitting the few-label probes. Each budget samples equal numbers of correct and incorrect examples from the source train+validation pool and uses identical sampled positions for both observers. The standardizer and probe are fitted on those sampled examples; Full uses the entire eligible source pool. The stated 4/8/16 budgets count labels used in this fit. Source-side configuration assessment uses the labeled source pool and is separate from this fit budget.

F.3 Baselines and capacity matching

Query-side baselines comprise class prior, query length, TF–IDF, query-only last-state and flattened probes, a length-matched ordinary filler, Difference-of-Means directions, FacLens (Wang et al., 2025), and Internal Confidence (Chen et al., 2026). Internal Confidence is reported both with its published training-free aggregation and with supervised logistic regression over its full layer–token $P(\mathrm{Yes})$ grid, giving it the same labels and classifier family as ThinkUQ. The answer-level Choice Summary uses permutation-invariant statistics of the choice distribution and is explicitly identified as an upper reference. All matched representation probes use the same source examples, preprocessing, layer candidate set, regularization grid, validation objective, and label budget unless an ablation states otherwise.

Table 16 summarizes representative answer-aware references. These methods receive information unavailable to a pre-generation observer, so they define a richer information regime rather than matched competitors. ThinkUQ remains competitive with answer-choice likelihood on MMLU-Pro and with length-normalized likelihood, a post-answer hidden state, and five-sample semantic entropy on TriviaQA. The larger gap on ARC illustrates the value of answer logits when the constrained policy is already highly accurate.

Table 16: Representative answer-aware references. Values are AUROC. Query and THINKUQ are pre-generation; the final column observes answer logits, a visible answer, or five sampled answers.

Task	Answer-aware reference	Query-end	THINKUQ	Reference
MMLU-Pro	Selected-choice likelihood	.6524	.7249	.7404
ARC-Challenge	Selected-choice likelihood	.6460	.6939	.8750
GPQA-Diamond	Selected-choice likelihood	.5019	.4463	.5224
TriviaQA	Length-normalized likelihood	.7943	.8238	.7931
TriviaQA	Post-answer hidden state	.7470	.7698	.8357
TriviaQA	Semantic entropy, $K = 5$	.6960	.7501	.7642

F.4 SELECTION, METRICS, AND UNCERTAINTY

We report AUROC, AURC, and NLL; AUROC measures ranking under different base accuracies and AURC measures selective deferral. For ten-fold transfers, the point estimate is the mean of the ten fold-specific AUROCs, and each paired bootstrap draw uses the same resampled indices for both observers. Ex jointly resamples frozen source folds and correctness-stratified target examples. Gp jointly resamples folds and the groups recorded in the benchmark: MMLU-Pro subjects, TriviaQA question sources, MATH subjects, and PopQA relations. GSM8K has one target group, so its Gp interval varies the source folds while retaining the complete target pool. HumanEval uses individual problem IDs as groups. A direction is resolved when both 95% percentile intervals exclude zero with the same sign; otherwise it is unresolved.

Bootstrap replication counts are collected here for reference. The primary comparison, cross-scale transfers and family synthesis, fixed-layer AUROC, and probability-shift analyses use 5,000 draws. The primary frozen-fold stability intervals use 100,000 draws. The ordinary-suffix, official-template, soft-prefix, residual-patching, prompt-surface, normalized-L2, margin, prequential-code, few-label, free-form-likelihood, token-response, attention-routing, and training-trajectory analyses use 2,000 draws. Each analysis retains its specified resampling units; prequential coding resamples source examples or source groups, and training checkpoints use paired correctness-stratified target examples. These intervals condition on the fitted probes and recorded model/training seeds and quantify variation from the resampled data or folds.

G EXACT INFORMATION BOUNDARY AND SERIALIZATION

The answer pass produces correctness labels from the unmodified query under the fixed task-specific answer policy. Each observer independently processes that query and its fixed suffix, with readout newlines supplied by teacher-forced prefill and zero decoded tokens. Source-fitted probes and preprocessing remain frozen during target evaluation. This common computation separates the information used to produce an answer from the state used to predict its correctness.

For the primary Qwen3-4B MMLU common-readout protocol, the query suffix is the shared newline token $[198]$ and the empty suffix is

$$[151667,\ 198,\ 198,\ 151668,\ 271,\ 198]. \tag{12}$$

The first five tokens are the piece-wise encoded empty reasoning block and the final token is the identical readout appended to every observer condition; the matched filler replaces those five block tokens by five period tokens. Canonical-template panels instead use the official four-token sequence $[151667, 271, 151668, 271]$. Both serializations decode to the model's native empty reasoning boundary, are fixed across queries, and decode no output token. Tagged-CoT training records use the official structured `reasoning_content` serialization.

H ADDITIONAL STATISTICAL RESULTS

Table 17 reports paired AUROC differences for the primary 3,000/1,000 study. Fold intervals summarize stability across the ten frozen source-subject partitions; hierarchical intervals combine that partition variation with target-example or subject resampling. Appendix F gives the resampling definitions and replication counts.

Table 17: Paired AUROC differences: THINKUQ minus comparator.

Comparator	MMLU Δ [fold CI]	MMLU-Pro Δ [fold CI]	Ex CI	Gp CI
Query-end state	$+.0995$ $[+.0802, +.1189]$	$+.0725$ $[+.0659, +.0809]$	$[+.0441, +.0999]$	$[+.0511, +.0929]$
Matched filler	$+.0631$ $[+.0510, +.0757]$	$+.0661$ $[+.0580, +.0761]$	$[+.0371, +.0963]$	$[+.0314, +.1010]$

Table 18: Additional validity checks. These analyses test whether the main effect can be reduced to text difficulty, split leakage, or source–target duplication.

Analysis	Coverage	Result
Shared/fixed layers	11 model–task points	The TriviaQA gain survives a shared source-selected layer at 1.7B ($+.0570$) and 14B ($+.0354$), with both example and group CIs above zero; the 14B result also survives a globally fixed depth.
Text difficulty	Length and TF–IDF controls	Shared-layer THINKUQ exceeds TF–IDF in 10/11 point estimates and is significantly better under both units at five points. The boundary gain and internal–text gap are both positive at 1.7B and 14B TriviaQA.
Split/label shuffle	11 points, 44 null checks	All 11 split/leakage checks pass. 43/44 shuffled-label 95% intervals contain .5; 2.2 exclusions are expected across 44 unadjusted 95% intervals. Six of 11 points satisfy the additional compound criterion combining a mean-AUROC band around .5 with interval coverage; five do not.
Contamination and deduplication	Six transfer datasets	Exact and lexical matching identifies five overlapping target records among 13,491; removing them leaves 99.963% of the target union. GPQA-Diamond, GSM8K→MATH, MBPP→HumanEval, and PopQA contain no detected cross-split match.

I ADDITIONAL SHORTCUT, TRANSFER, AND UTILITY ANALYSES

Table 18 summarizes additional validity checks. "Positive" means agreement between paired example and group bootstrap at the prespecified 95% threshold.

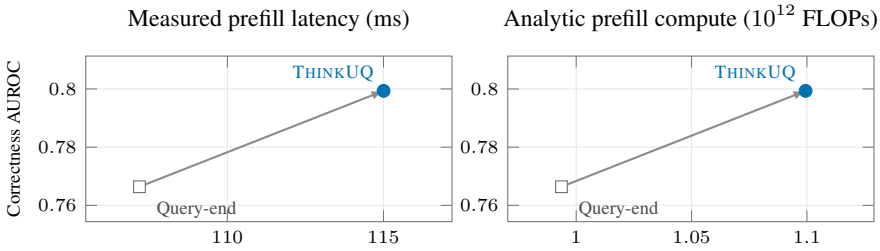

Figure 5: Measured observer quality–cost trade-off on Qwen3-14B TriviaQA. Answer accuracy is fixed at .4483. Query-end/ThinkUQ correctness AUROC is .7664/.7993 ($\Delta = +.0330$; paired split+example/group CIs $[+.0125, +.0539]/[+.0010, +.0597]$). Mean prefill latency is $107.18/115.00$ ms, while analytic FLOPs increase by 10.65%. Both modes are Pareto-efficient: the boundary buys higher ranking quality at higher measured cost.

Table 19: Prompt-surface robustness on Qwen3-4B MMLU→MMLU-Pro ($400/200$ source/target; target accuracy .4450). Intervals are paired hierarchical fold+example (Ex) and fold+group (Gp) 95% CIs. Canonical is common to both regimes.

Regime	Surface	Query-end	ThinkUQ	Δ	Ex 95% CI	Gp 95% CI
Common	Canonical	.6061	.6867	$+.0807$	$[+.0196, +.1414]$	$[+.0012, +.1615]$
Matched surface	Paraphrase	.5772	.6737	$+.0965$	$[+.0249, +.1735]$	$[+.0085, +.1920]$
Matched surface	Option permutation	.6050	.6441	$+.0392$	$[-.0417, +.1195]$	$[-.0810, +.1479]$
Canonical transfer	Paraphrase	.5817	.6598	$+.0782$	$[+.0003, +.1473]$	$[-.0106, +.1609]$
Canonical transfer	Option permutation	.6185	.6835	$+.0650$	$[-.0012, +.1333]$	$[-.0187, +.1549]$
Perturbation-minus-canonical difference-in-differences						
Matched surface	Paraphrase	–	–	$+.0158$	$[-.0422, +.0758]$	$[-.0528, +.0871]$
Matched surface	Option permutation	–	–	$-.0415$	$[-.1104, +.0265]$	$[-.1221, +.0193]$
Canonical transfer	Paraphrase	–	–	$-.0025$	$[-.0568, +.0532]$	$[-.0582, +.0554]$
Canonical transfer	Option permutation	–	–	$-.0156$	$[-.0629, +.0302]$	$[-.0544, +.0219]$

Two-sided paired-bootstrap p-values for the four lower-panel rows (Ex/Gp) are .578/.733, .222/.196, .934/.862, and .497/.428, respectively; every contrast is unresolved under both units.