CHAIN-OF-THOUGHT RELOCATES THE FACTUALITY SIGNAL: A LOCKED-PROTOCOL AUDIT AND CLAIM-PAIRED RECOVERY OF INTERNAL PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Internal-state factuality probes do not lose their signal under chain-of-thought: the signal moves. Across four model families, under a single locked protocol that varies labeling budget, training style, and feature pipeline orthogonally, we make three measurements. First, the direct-to-CoT transfer loss is small relative to protocol choices: the feature pipeline alone shifts CoT-side scores by an order of magnitude more than the style change itself, and target-style labels close most of the residual. Second, the residual is structured internal geometry: a claim-paired affine map recovers it, shuffle-paired controls collapse to chance, generic unsupervised alignment fails on every family, and label-free logit distillation from the same pairs nearly matches the full map – recovery is a label-transfer phenomenon that runs through the claim-paired internal geometry rather than through surface form. Third, natural false claims are measurable without synthetic corruption, and the labels themselves are auditable: adjudicating every generation shows that automatic labels conflate asserted errors with unasserted truncations, and a strict audit retracts two of our own label regimes as artifacts – yet the detection signal survives label cleanup. Protocol and labels determine what a factuality probe measures; the audit itself is the transferable finding, with claim localization and corpus-specific generalization the open frontiers.

1 INTRODUCTION

Reasoning large language models state more facts per answer than non-reasoning models, and they state more false ones (Chen et al., 2026). An internal-state probe reads a claim's truth from hidden activations without any external reference (Azaria & Mitchell, 2023), which makes it the natural instrument for monitoring generation as it happens. The obstacle reported in the literature is that chain-of-thought obscures the very cue the probe reads (Cheng et al., 2025). This paper measures that obstacle and finds it smaller and differently shaped than reported: *the factuality cue survives CoT and moves to a position that target-style labels and a claim-paired map both reach.*

The starting puzzle is a number from our own earlier experiments, a direct-only arm at 0.80 against a mixed arm at 0.999. That configuration differed between arms in three things at once: supervision budget, generation length, and the rule that locates the asserted claim. Once budget is matched and spans are located consistently, that 0.195 arm gap does not survive: the mixed arm lands within -0.010 to $+0.028$ AUROC of the direct-only arm, and the residual transfer loss is 0.014–0.040 (Section 5). What remains is a real effect, and its size is best judged against the feature pipeline: swapping raw concatenated states for a PCA projection moves the same probe by 0.13–0.28 AUROC, 3–$20\times$ more than the cross-style transfer it is usually compared against. Auditing the protocol is therefore not bookkeeping; it determines which effect is worth explaining.

We organize the paper around three measurements.

Measurement 1: the transfer loss, with the confounds separated. Under one locked protocol – identical subject split, identical test items, five probe seeds per cell, budgets counted in absolute labeled spans – we cross budget, training style, and feature pipeline. The direct-only probe loses 0.014–0.040 AUROC when it scores CoT claims under the raw pipeline (0.13–0.21 under PCA-256:

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.

the pipeline amplifies the style cost as well as shifting the level). Target-style labels close the residual, and the target-only arm's gain grows with budget: $+0.004$ at 60 labeled spans, $+0.034$ at 960 (the mixed arm trails direct-only at B=60). Under the PCA pipeline the same labels add $+0.19$ at 60 and $+0.13$ at 960.

Measurement 2: the geometry of what moved. A claim-paired affine map – fitted on the same claim stated in both styles, using no factuality label – restores the frozen probe on Qwen and Llama. Three controls locate the effect in the representation: shuffle-paired pairing collapses to 0.46–0.52; pair-free generic alignment (mean-shift, CORAL, Procrustes) never recovers the paired gain on any family (0.37–0.95); a text-only paired model, granted the same budget and calibration, reaches 0.544 where the hidden-state probe reaches 0.958. Per-family selection among the cached layers changes every family by at most 0.008. Gemma does not recover under any of these conditions: the geometry, under this map protocol, is a property of the model family and a target for mechanistic work.

Measurement 3: where natural errors actually are, and whether the labels can be trusted. Monitoring is only useful if it catches errors the model makes on its own, so we measure error populations without injecting any – and we audit the resulting automatic labels, including our own. On a hard PopQA slice (Mallen et al., 2023), the model's own generations supply 418 false claims on Qwen and 385 on Mistral; a probe trained and evaluated on those natural labels reaches 0.988 and 0.998 AUROC, agrees with independent relabeling at κ=0.96, and transfers to TriviaQA (Joshi et al., 2017) at 0.937 and back at 0.966. In the synthetic single-hop pool, every one of 493 confident in-KB assertions is true: the model's errors there take the form of truncation, hedging, or fabricated values outside the knowledge base, not confident misstatement. On PrOntoQA 2-hop (Saparov & He, 2023), our automatic symbolic labeling initially reported 20.6% false intermediate steps with a 0.982 step-level probe; the surface audit below flagged those labels, root-causing a parser artifact whose correction leaves a true step error rate of 0.5% (5 of $1{,}064$ verifiable steps, each independently re-derived): natural step errors on structured domains are rare enough that sampling them at scale needs harder domains or injected distractors. The audit – a six-feature surface classifier trained to predict the label – is cheap, label-free, and catches label corruptions that cross-validation, held-out agreement, and probe AUROC all miss.

Contributions.

1. *The CoT collapse we reported is a protocol effect in our setup.* With budget, style, and pipeline varied orthogonally under one protocol, transfer loss is 0.014–0.040 AUROC and pipeline choice dominates it $19.7\times$ at B=60 and $3.2\times$ at B=960; target-style supervision closes the residual, with adaptation gains reported per arm (target-only $+0.004$ to $+0.034$; the mixed arm trails direct-only at B=60).

2. *The residual is structured, internal, and family-dependent.* A claim-paired affine map recovers it on Qwen and Llama; shuffle, generic alignment, text-only, and layer-choice controls all fail to; a same-pairs logit distillation matches the full map within 0.005, locating the active ingredient in label-level transfer rather than full-geometry transport; Gemma does not recover.

3. *Natural false claims are measurable, and the labels are audited.* On hard natural-error slices, alias-label probes reach 0.988–0.998 and validate at κ=0.96; strict adjudication of every generation shows the alias labels conflate asserted errors with truncated non-assertions (27.5% of auto-FALSE items assert a wrong answer, 24.7% assertion coverage; assertion-only probe 0.82–0.99 mean 0.93 with question-cluster CIs). A six-feature surface screen reads 0.54/0.58 (z=1.4/3.9 against label permutation) on those labels and 0.996 on our own mis-labeled PrOntoQA step pool, whose true error rate is 0.5% – the screen retracts a headline result of our own and is proposed as standard practice for automatic natural-error labeling.

Roadmap. Section 2–Section 4 give background and setup; Section 5 the three measurements, detector comparisons, and ablations; Section 6 and Section 7 close.

2 Related Work

Prior work sorts along one axis that matters for deployment: is the method *reference-free*, is it *training-free* with respect to the monitored model, and does it *hold up under CoT*? Our probe is reference-free and leaves the model unchanged, and its central result reframes the third question: the CoT shift of the factuality cue is a *family-dependent recoverable readout transfer* – a few hundred labels, or claim pairs alone, restore the readout.

Internal-state factuality probes. Azaria & Mitchell (Azaria & Mitchell, 2023) showed that a lightweight neural classifier on hidden states predicts whether the model's own statement is true, seeding the reference-free detection line. Extensions cover temporally consistent factuality probing (Bajpai et al., 2024), fact-checking against generated rationales, and richer probe architectures. These methods train and evaluate on *direct* answers, so their behavior under CoT is the open question this paper measures: under a locked protocol the raw-pipeline cross-style transfer loss is 0.014–0.040 AUROC (0.13–0.21 under PCA-256) (Section 5), and the remaining question is how the readout reaches the shifted cue.

CoT changes the detector's input. A pointed literature reports that reasoning changes the internal signal detectors rely on: CoT obscures the hallucination cues that direct-answer probes exploit (Cheng et al., 2025), spanning four datasets, four models, and eight *training-free* detector families; it excludes supervised probes by design, on generalization grounds. This paper measures the complementary supervised slice and separates the protocol confounds from the style effect: budget, pipeline, and span rule jointly manufacture the large apparent drop, and what remains after matching is 3–$20\times$ smaller than the feature-pipeline effect (0.014–0.040 vs. 0.13–0.28) and is closed by target-style labels or a claim-paired map (Sections 5.2 and 5.3).

CoT-aware shaping, uncertainty heads, and stronger readouts. A separate line changes the *detector* rather than the measurement protocol: ARS (Zhang et al., 2026) shapes trace-boundary representations by answer agreement; pre-trained uncertainty heads (Shelmanov et al., 2025) supervise modules on attention features; token-/layer-selective probes (Meyoyan & Del Corro, 2026) pick informative tokens and layers from the hidden-state tensor; Truthfulness Separator Vectors (Park et al., 2025) steer latents; semantic-entropy probes (Kossen et al., 2024) approximate semantic entropy from one generation's states; FacLens (Wang et al., 2025) predicts non-factuality ante hoc; INTRA (Vazhentsev et al., 2026) adds multi-layer interaction features, HARP (Hu et al., 2026) projects hidden states onto an unembedding-derived reasoning subspace, and DRIFT (Bhatnagar et al., 2026) probes intermediate-layer confidence alongside generation. The measurements here bound what readout engineering can buy: architecture is not the active ingredient – ICR-style dynamic probes and INTRA-style interaction features transfer as well as the plain concatenated probe, per-family layer re-selection moves scores by at most 0.008, and TSV-style steering fails outright (Section 5.7, Appendix A.7, A.18) – while labels and pairs move the score. The claim-paired map is a linear alignment akin to model-stitching alignments such as HELIX (Gorbett & Jana, 2026), but *within* one model across styles and *claim-paired*: generic alignments fail without claim correspondence (Appendix A.10).

Benchmark audits and unified detector evaluations. A growing line audits hallucination-detection benchmarks rather than proposing detectors: PARALLAX (Hussain & Kantarcioglu, 2026) shows corpora that embed the answer in the prompt inflate detectors to near-perfect AUROC; OpenHalDet (Li et al., 2026) standardizes evaluation across access levels; a factorised study (Srey et al., 2026) finds raw hidden states hard to beat in-domain while structured features transfer better under shift. This paper adds the two confounds the audits had not isolated – supervision budget and feature pipeline – plus the span-rule confound, and shows they jointly manufacture the reported collapse (Section 5.2). At claim granularity, HalluEntity (Yeh et al., 2025) shows entity-level detection remains unsolved for uncertainty methods; our localization cost of 0.217 AUROC (single-seed oracle-span slice) pins the same frontier from the internal-state side (Section 5.5).

Inference-time decoding and intervention. DoLa (Chuang et al., 2024) contrasts early/late-layer logits; SelfCheckGPT (Manakul et al., 2023) scores sampling consistency. Both are training-free and reference-free but operate on output tokens or sampled generations rather than on a calibrated monitor

of the claim span. The probe is complementary; on identical adjudicated items it reads 0.82–0.99 against p(True) at chance and calibrated span-entropy at 0.44–0.88 (Section 5.6).

Factuality via RL and reasoning-model hallucination. RL fine-tuning with factuality rewards (Chen et al., 2026) improves truthfulness by changing the model at the cost of a reward model or retriever. This paper's program is the other half: leave the model alone and monitor it. The finding that the factuality signal lives in middle layers (Geva et al., 2021; Azaria & Mitchell, 2023) motivates the layer set (Appendix A.11).

3 Background and Problem Setup

A *claim* is a span of tokens the model asserts, such as an object entity. We extract a fixed-dimensional vector $\mathbf{h}_S$ by mean-pooling hidden states at a chosen layer set $\mathcal{L}$ and concatenating (Equation (2) in the appendix). A *probe* maps $\mathbf{h}_S$ to a truth score $s(S) \in [0, 1]$. The probe is fitted on short direct answers drawn from $\mathcal{P}_{\text{direct}}$ and is then asked to score claims produced after multi-step CoT, drawn from $\mathcal{P}_{\text{reason}}$.

Writing ϕ_{dir} and ϕ_{reas} for the two distributions of $\mathbf{h}_S$, the quantity of interest is the drop in s quality when the probe crosses from ϕ_{dir} to ϕ_{reas}. Two interventions close it. Target-style supervision adds labeled spans from $\mathcal{P}_{\text{reason}}$ to the training set, which moves the decision boundary. A claim-paired affine map instead applies $T(h) = Wh + b$ fitted on pairs $\{(h^{(i)}_{\text{reas}}, h^{(i)}_{\text{dir}})\}$ that state the same claim in both styles, leaving the probe fixed; because pairing uses surface claim identity alone, the map consumes no factuality label. Both are defined in Section 4.

The difference between them is what each one assumes. Supervision assumes the reasoning distribution carries learnable structure; the map assumes the two distributions are related by an affine transformation conditioned on claim identity. Section 5 tests both and, more importantly, tests them against controls that share their budget and their calibration: a shuffle-paired map, pair-free alignment, a text-only paired model, and per-family layer selection.

4 Method: Two Interventions and Their Controls

Two interventions close the cross-style residual that Section 5 measures: target-style supervision, which moves the decision boundary, and a claim-paired affine map, which moves the representations. Both leave the monitored model untouched.

Probe. The paired-map experiments use a 2-layer MLP head (one 256-dim hidden layer with ReLU, $\mathbf{h}_S \mapsto \mathbb{R}$) on the pooled claim-span representation $\mathbf{h}_S$ from Equation (2), trained with binary cross-entropy (AdamW). The layer set $\mathcal{L} = \{4, 10, 18, 24\}$ is fixed across all families; the signal peaks in the middle layers (L15–L19) and any mid-to-late selection is near-optimal (Appendix A.11), and Section 5.3 shows per-family re-selection moves results by at most 0.008. The unified budget×style×pipeline factorial (Section 5.2) uses an L2 logistic head on the same cached features – the readout is an explicit experimental factor there: rerunning the key cells (raw/PCA-256 $\times$ direct/target, $B{\in}\{60, 240, 960\}$) with this MLP head reproduces every factorial conclusion (target-only beats direct-only in $6/6$ cells under both heads; raw transfer loss small and PCA loss large under both), with per-cell level deviations of -0.11 to $+0.02$ AUROC (Block 72b, Appendix A.5e).

Intervention 1: target-style supervision. The probe is trained on a mix of direct and reasoning-style renderings of the same underlying knowledge,

$$\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{direct}} \cup \mathcal{D}_{\text{reason}}, \qquad |\mathcal{D}_{\text{direct}}| : |\mathcal{D}_{\text{reason}}| = 1 : 1, \tag{1}$$

so the only change from a direct-only probe is exposure to the reasoning distribution $\mathcal{P}_{\text{reason}}$ at training time. Section 5.2 measures the adaptation arms separately as a function of budget: target-only gains $+0.004$ at 60 labeled spans and $+0.034$ at 960 on the raw pipeline (the mixed arm trails direct-only at $B{=}60$), and the same target-style labels add $+0.19$ down to $+0.13$ under PCA-256 – the gain scales with the residual it corrects. An ICR-style cross-layer dynamic probe transfers across styles as well as the static concatenated probe (Section 5.7), so the gain comes from the supervision, not from the readout.

Intervention 2: the claim-paired affine map. To measure *what kind* of transformation the shift is, we fit a ridge-regularized affine map $T : \mathbf{h}^{\text{reas}} \to \mathbf{h}^{\text{dir}}$ on *claim pairs* – the same (subject, relation, object) rendered in the direct and reasoning styles, matched by surface content, *no factuality label* – minimizing $\|T\mathbf{h}^{\text{reas}} - \mathbf{h}^{\text{dir}}\|_2^2$ with the penalty selected on a subject-disjoint held-out split. Applying T to reasoning-style states lets the frozen direct-only probe score them as if they were direct. If the shift were a global offset, a mean-shift would recover it; if it were absent, no map would be needed. Section 5.3 shows the paired map recovers $0.919 \to 0.983$ (Qwen) and $0.923 \to 0.974$ (Llama) while shuffle-paired, pair-free, text-only, and layer-reselection controls all fail, and a same-pairs logit distillation matches the full map: the recovery is a measurable, label-level cross-style readout transfer carried by the paired geometry, and the controls bound what kind of transformation it is not – a global offset, a generic prompt effect, or a text artifact. The finer decomposition is a finding of Section 5.3, not an assumption of this section.

Controls. Every intervention is tested against controls that share its budget and calibration: the shuffle-paired map (does correspondence carry the effect?), pair-free generic alignment (mean-shift, CORAL, PCA-Procrustes), the text-only paired probe (is the signal in the surface?), per-family layer selection (is it the readout?), and reference-free detectors on identical items (is the supervised probe adding anything?). The controls are the evidence; the interventions are what they license.

One head. A single recalibrated head covers both styles. A two-head design with an oracle style flag is unnecessary: at deployment budgets ($B \geq 480$), the mixed head lands within 0.003–0.008 of the direct-only head's own-distribution accuracy while also covering CoT (Appendix A.2).

5 Experiments

5.1 Setup

Task and data. Each controlled-KB example is a (subject, relation, object) triple; a TRUE claim asserts the gold object, a FALSE claim a same-relation distractor. The KB has 64 unique subjects and 125 unique (subject, relation) facts over 10 relations (82 per class in the train pool); training rows resample them, so labeled spans, not independent facts, are the counted resource: at $B{=}60$ the 30 spans per class are 30 distinct facts, but at $B{=}960$ the pool cap binds and each fact appears $\sim$6 times per class. The same claim is rendered in two styles – a short *direct* answer and a *reasoning* answer wrapping the claim in templated CoT – with identical claim-span markers, isolating templated-context change, a controlled slice of real CoT (real traces in Section 5.5). Train and test subjects never overlap (fact-disjoint split), ruling out memorization. All factorial cells (Section 5.2) share one subject split, one test set, and five probe seeds; budgets are absolute counts of labeled spans.

Models and probe. Four model families: Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct, Gemma-3-12B-Instruct, plus Qwen2.5-1.5B as a size diagnostic. The probe reads the pooled claim-span representation $\mathbf{h}_S$ (Section 3) from layers $\{4, 10, 18, 24\}$ with a 256-unit MLP head, one configuration selected on Qwen and applied *unchanged* to all families (Section 5.3: per-family layer re-selection moves results by at most 0.008).

Statistical unit. A cell mean is over probe seeds on a cached partition; significance statements are subject-level cluster bootstraps or seed-paired tests, named in place.

5.2 Measurement 1: the transfer loss with confounds separated

The factorial. We cross budget $B \in \{60, 120, 240, 480, 960\}$ total labeled spans $\times$ training style (direct-only / target-only / mixed 1:1) $\times$ feature pipeline (raw cross-layer vs. PCA-256 fit on the pooled training rows), one locked protocol, five probe seeds per cell (Table 1; full grid in Appendix A.5e).

Transfer loss: 0.014–0.040 AUROC under raw; 0.13–0.21 under PCA. The direct-trained probe scores CoT claims 0.014–0.040 AUROC below its own direct-style performance at every budget – under the raw pipeline; the same arm on PCA-256 features loses 0.13–0.21: the style cost is pipeline-dependent. The four-family matched-budget grid shows the same shape (0.92–0.95 cross-style at

Table 1: **Unified factorial under one locked protocol (Qwen-7B, 5 seeds, AUROC on the reasoning-style test). Top:** raw-pipeline transfer loss (direct-only in-style minus cross-style) is 0.014–0.040 at every budget – and 0.13–0.21 under PCA-256, 4–10× larger. **Bottom:** the pipeline main effect dominates the raw cross-style effect 19.7× at B=60, 3.2× at B=960; target-style labels close the residual (details in text).

Raw pipeline: direct-only arm, in-style vs. cross-style					
Budget	60	120	240	480	960
direct-style test	0.943	0.968	0.979	0.984	0.990
CoT-style test	0.929	0.932	0.946	0.949	0.950
transfer loss (raw)	0.014	0.036	0.033	0.035	0.040
transfer loss (PCA-256)	0.175	0.213	0.189	0.136	0.132

Cross-style AUROC by training style and pipeline						
	direct-only		target-only		mixed	
	raw	PCA	raw	PCA	raw	PCA
B=60	0.929	0.653	0.934	0.846	0.919	0.740
B=240	0.946	0.735	0.965	0.885	0.947	0.881
B=960	0.950	0.822	0.984	0.948	0.978	0.949

B=1920 on Qwen, Mistral, and Llama; Gemma 0.80, Appendix A.5h). The early 0.804-vs-0.999 collapse (Section 1) came from changing budget, span rule, and pipeline at once; under the locked protocol it is a 3–4-point effect.

Pipeline dominates style, by a budget-dependent ratio. At identical budget and split, swapping raw for PCA-256 moves the direct-only probe by 0.128–0.276 AUROC on the CoT test, against a raw cross-style loss of 0.014–0.040: any evaluation that changes pipeline and style together measures the pipeline, not the style.

Target-style labels close the residual, with gains that scale. Target-only: +0.004 at B=60, +0.034 at B=960 over direct-only (raw). Mixed (same total budget): −0.010 at B=60 – below direct-only – and +0.028 at B=960: at low budget, spending half the labels on direct-style data is a net loss. Under PCA-256 the same labels add +0.19/+0.13 – the gain grows exactly where the residual is.

5.3 Measurement 2: The geometry of the residual

In-style separability survives CoT. Factuality stays linearly separable *within* the CoT distribution (in-style AUROC 0.89–0.99, Figure 1a) – the signal is present after CoT; the direct-trained probe reads it from the wrong place. Direct↔CoT CKA falls to 0.46 at the top layer (panel c): the two styles occupy different geometry (Appendix A.10).

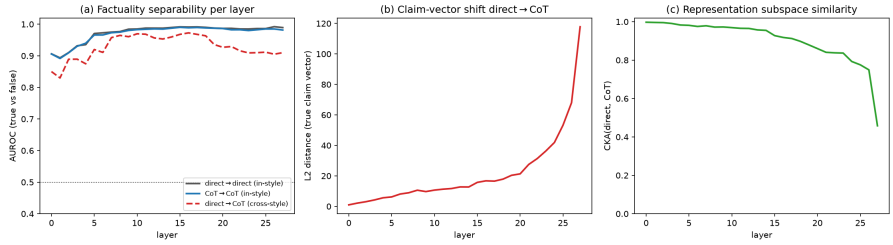


Figure 1: **The signal survives CoT and moves.** (a) Per-layer factuality separability: in-style direct (gray) and in-style CoT (blue) both stay high, while the cross-style direct→CoT probe (red, dashed) drops – the cue is relocated, not lost. (b) Claim-vector L2 distance grows with depth. (c) CKA falls to 0.46 at the top layer.

Table 2: **Claim-paired map recovers the frozen direct-only probe.** Map $T : \mathbf{h}^{\text{reas}} \to \mathbf{h}^{\text{dir}}$ fit on claim pairs with no factuality label; controlled-KB rows are mean±std over 3 subject-disjoint partitions; "CoT ref." is a truth-supervised CoT probe; PopQA row uses 5 partition seeds and gold-conditioned pairing (Appendix A.8) – its same-assertion analogue (no answer key) reaches 0.953 ± 0.010 (Section 5.3). Shuffle pairs → chance on every KB family (on PopQA the shuffle mean is chance-level, 0.527 ± 0.049, with wider per-seed spread): within the tested map family, paired fitting is the active ingredient; the label-preserving decomposition is Block 71.

Family	Direct-only	Aligned	CoT ref.	Shuffle	Recovery	Sig.
Qwen-7B	0.919 ± 0.012	$\mathbf{0.983}\pm0.000$	0.972 ± 0.003	0.463 ± 0.045	$+0.064\pm0.012$	$3/3$
Llama-3.1-8B	0.923 ± 0.005	$\mathbf{0.974}\pm0.001$	0.949 ± 0.013	0.516 ± 0.087	$+0.051\pm0.006$	$3/3$
Mistral-7B	0.948 ± 0.007	0.958 ± 0.001	$\mathbf{0.957}\pm0.006$	0.465 ± 0.089	$+0.010\pm0.007$	$0/3$
Gemma-3-12B	0.786 ± 0.030	0.777 ± 0.021	$\mathbf{0.878}\pm0.013$	0.477 ± 0.010	-0.010 ± 0.022	$0/3$
PopQA (real)	0.718 ± 0.027	$\mathbf{0.840}\pm0.038$	—	0.535 ± 0.207	$+0.122\pm0.042$	$4/5$

A claim-paired affine map recovers the frozen probe. We fit a ridge-regularized affine map $T : \mathbf{h}^{\text{reas}} \to \mathbf{h}^{\text{dir}}$ on claim pairs – the same (subject, relation, object) rendered in both styles, matched by surface content, *no factuality label* – and apply the *frozen* direct-only probe to mapped CoT states. On held-out facts the probe recovers from 0.919 to 0.983 on Qwen-7B ($+0.064$, $3/3$ seeds) and from 0.923 to 0.974 on Llama-3.1-8B ($+0.051$, $3/3$), reaching the truth-supervised CoT reference (Table 2); over 5 seeds the paired gap (aligned − CoT reference) is positive on both families (Qwen $+0.009$, p=0.026; Llama $+0.025$, p=0.006; Holm: Llama p_{adj}=0.018, Qwen 0.053).

Four controls locate the effect in the representation.

1. *Shuffle-paired control:* scrambling which direct vector each CoT vector is mapped to collapses recovery to chance on every KB family (0.46–0.52, Table 2; PopQA's same-assertion shuffle sits at 0.527 ± 0.049): pair-free fitting of the same family fails – within the tested map family, paired fitting is the active ingredient. *Label-preserving shuffle:* permuting partners *within* each truth class reaches 0.972 (Qwen) and 0.957 (Llama) against the paired $0.982/0.974$ (Block 71, Appendix A.10): the class-conditional component delivers $+0.057/+0.031$ over no map; the residual exact-correspondence increment is $+0.010/+0.017$ as a difference of the displayed estimates, and the paired subject-cluster bootstrap of the same increment reads $+0.015/+0.013$ with CIs straddling zero ($[-0.004, +0.050]/[-0.003, +0.035]$, Block 74).
2. *Logit-distillation control (Block 74):* a same-pairs ridge from CoT feature to the frozen probe's *logit* on the paired direct feature – scalar target, no labels – reaches $0.980/0.969$ against the full map's $0.982/0.974$ (within 0.005): the paired fit transfers, to first order, the truth-relevant score – label-level distillation, not full-geometry transport.
3. *Pair-free generic alignment:* mean-shift, CORAL, and PCA-Procrustes fit on the same states without pairs all fail on every family (Qwen: $0.925/0.842/0.652$ vs. 0.983 paired; Appendix A.10); what the paired fit supplies and the pair-free fits lack is the class-conditional structure identified above.
4. *Text-only and layer controls:* a TF-IDF probe on the identical texts, budget, and 300 test items reaches 0.544 vs. 0.958 for the hidden-state probe (Appendix A.7); per-family layer re-selection over all 15 cached subsets moves test AUROC by at most 0.008 (full-depth selection untested).

The map's structure. Paired Procrustes ($0.956/0.944$) and PLS-64 ($0.962/0.952$) also recover but trail the affine ridge; CCA-64 fails ($0.46/0.44$). The map is one-directional (effective rank $217/256$, reverse transfer 0.49, cycle error 0.47 rel. RMSE) – linear recoverability, not invertibility. The same frozen T recovers an independently trained linear probe ($0.827 \to 0.984$ Qwen; $0.942 \to 0.976$ Llama) and survives object-disjoint testing (Appendix A.14–A.15).

Family dependence is real. Gemma-3-12B carries a healthy in-style CoT signal (0.878 ± 0.013) that no tested map recovers: ridge, Procrustes, PLS, CCA, an MLP map, k-NN, and all 15 cached-layer subsets fail; Mistral recovers marginally (baseline near-saturated, 0.948). Rerunning the protocol

unchanged on a second checkpoint per family (Block 70) replicates the split: Qwen2.5-1.5B and Ministral-8B recover, Gemma-3-4B fails, and the one non-recovering non-Gemma is Llama-3.2-3B – the only checkpoint built by pruning-and-distilling rather than training from scratch; across eight checkpoints the split tracks pretraining lineage (a correlation, not causal; Appendix A.7).

Cost of pairing. Pairs are free in the templated setting. On PopQA, two pairing rules at identical pipeline: *gold-conditioned* (each CoT row maps to its question's direct-*TRUE* rendering; gold enters through the target on wrong-answer rows) and *same-assertion* (the direct side renders the model's *asserted* object, no answer key; Block 75b). Same-assertion reaches 0.953 ± 0.010 vs. gold-conditioned 0.840 ± 0.038 (5 partition seeds, Table-2 protocol) – dropping the answer key *raises* the score: gold-conditioning breaks the pair exactly on wrong-answer rows. Shuffle 0.527 ± 0.049; label-free logit distillation 0.950 ± 0.004; recovery $5/5$ significant vs. unadapted 0.718 (the single-seed gold 0.954 was a pooling artifact, Appendix A.8): **the real-corpus recovery does not depend on the answer key**. 200 questions yield $192/194$ localizable train/test pairs (96%/97%; Appendix A.8).

5.4 Measurement 3: natural errors and where they live

Monitoring earns its cost on the model's own errors: this subsection measures natural-error populations, all labeled automatically, none injected.

Hard natural-error slice (PopQA). On the hard/long-tail slice of PopQA, the model's own greedy generations supply 418 false claims on Qwen-7B and 385 on Mistral across 3 seeds (Table 3; alias labels, re-adjudicated below). A probe trained and evaluated on these natural labels reaches 0.988 ± 0.002 / 0.998 ± 0.001 AUROC; p(True) on the same items only 0.519–0.617. Natural base rates (0.58–0.78) are reported with every number.

Table 3: **Natural-error benchmark under alias labels** (3 seeds; probes trained on the model's own false claims; κ=0.96 vs. blind relabel on the 48-item adjudicable sample; p(True) on identical items). The alias labels are re-adjudicated in §5.4 (27% of auto-FALSE carry asserted errors): read as alias-population scores.

Family	natural false (test)	AUROC	AUPRC	FPR@95	p(True)
Qwen-7B	129–151	0.988 ± 0.002	0.996	0.04–0.12	0.519–0.564
Mistral-7B	115–142	0.998 ± 0.001	0.999	0.00	0.569–0.617

Label validity and three-way adjudication. A blind relabel of 103 items agrees with the alias labels on 97.9% of the 48 adjudicable (κ=0.96; the 55 excluded truncations are label-blind, $24/24$). Adjudicating all 1,185 kept test generations by terminal-assertion form finds 27.5% of auto-FALSE items assert a wrong answer: one quarter confident errors, three quarters unfinished text (coverage 24.7%: 56 TRUE-assert / 237 FALSE-assert; Block 73). Retraining on adjudicated assertions only gives 0.82–0.99 AUROC (mean 0.932 vs. 0.993 alias; question-cluster CIs in Appendix A.7); the frozen alias probe on that subset reads 0.97: the headline is an alias-population score whose assertion-only analogue is 0.93. An independent zero-shot LLM judge (Block 76, A.7) confirms the cue survives semantic labels (0.98/0.99 judge-labeled retrain on the two large-n runs; the small-n third run reads chance) but shows the regex overcounts false assertions (84% of regex-FALSE are truncated). The PopQA-trained probe also transfers zero-shot to TriviaQA at 0.937 (in-corpus 0.967; back-transfer 0.966) under the same alias rules.

Label audits: a screen that retracted our own result. One cheap control checks every automatic label: a logistic classifier on six surface features trained by 5-fold CV – a leakage *screen* read against a label-permutation null. The PopQA natural labels read 0.536/0.583 (Qwen/Mistral; z=1.4/3.9): limited surface predictability, not an upper bound on all shortcuts. Our own PrOntoQA step labels read 0.996: a "Therefore/Since/so" prefix parses as nested and never matches a closure atom, so every prefixed assertion was labeled false; the true step error rate is $5/1{,}064$, and the 0.982 step probe built on those labels is retracted as format-reading (Appendix A.7).

5.5 Real-corpus generalization

The synthetic-trained probe transfers to unseen PopQA (Mallen et al., 2023) at 0.772 ± 0.037 (Qwen) / 0.915 ± 0.052 (Mistral); 200 real labels lift this to 0.974 ± 0.004 / 0.996 ± 0.003 (Appendix A.8). On TriviaQA (Joshi et al., 2017), real-CoT recalibration lifts AUROC $0.823 \rightarrow 0.952$; on TruthfulQA (Lin et al., 2022), $0.385 \rightarrow 0.573$; on DeepSeek-R1-Distill-Qwen-7B native traces, 136 target-style labels reach 0.935 vs. 0.827 for $1{,}200$ direct. These rows use gold-answer span localization; automated localization costs 0.217 AUROC (Appendix A.12).

5.6 Detector comparisons as audit history

An earlier draft ranked the probe against reference-free detectors on spontaneous-CoT items; those numbers predate the label audit and were scored against the retracted proxy labels. We retain them in Appendix A.7 as measurement history and withdraw any current-performance ranking. The clean same-instance comparison now exists (Block 77, all detectors on identical adjudicated test items): retrained probe 0.82–0.99, frozen alias probe 0.96–1.00 (mean 0.97), p(True) 0.43–0.53 (chance), calibrated span-entropy 0.44–0.88 (Appendix A.7).

5.7 Ablations

Architecture is not the active ingredient: an ICR-style cross-layer dynamic probe and its static union transfer within 0.960–0.966 (Appendix A.18); at equal budget $2400{:}0$ direct leaves AUROC 0.943 vs. $1920{:}480$'s 0.982; the probe is natively calibrated (ECE 0.064, isotonic degrades ranking); one-dimensional recalibration and unsupervised alignment fail where the paired ridge recovers; CPU-only shortcuts sit at 0.50–0.53 (Appendix A.11). The PCA-256 arm fits its projection on $8{,}000$ unlabeled rows of both styles, so "direct-only" means direct *labels* only, not no access to target-style features.

6 Discussion

(i) In the simplified variants tested, labels beat architecture: every architectural variation leaves the residual, every label-based intervention removes it (Section 5.7). The practical lever is therefore the pairing structure of the supervision – matching spans and styles between training and deployment – rather than a better probe: it closes most of the gap at unchanged probe capacity. (ii) Within the tested layers, pipelines, and checkpoints, recovery is family-dependent, not probe-dependent: Qwen and Llama recover, Gemma fails, second checkpoints replicate the split, and a same-pairs logit distillation matches the full map – label transfer, carried by the claim-paired geometry (Section 5.3). A deployment should carry over the paired-fit recipe with its shuffle control and re-fit on its own checkpoint, rather than transplant a specific map. (iii) Labels must be audited beyond agreement statistics: the surface screen caught a parser artifact cross-validation had passed; three-way adjudication separated asserted errors from truncations (Section 5.4). High labeler agreement is not label validity – agreement with the automatic labeler coexisted with a majority of auto-FALSE items carrying no asserted answer.

7 Conclusion

We audited internal factuality probing across the chain-of-thought shift. The factuality cue survives chain-of-thought and moves (0.014–0.040 AUROC across styles vs. 0.13–0.28 for pipeline); target-style labels close it, and a claim-paired map recovers it – pair-free controls fail, label-free distillation matches, and recovery needs no answer key. Natural false claims are audited to the assertion level by form (27.5% of auto-FALSE; adjudicated-only mean 0.93; the cue survives independent LLM relabeling). For practitioners, the recipe is concrete: lock the protocol, train on target-style labels over paired claim spans, apply the same-pairs map, and audit the labels before trusting any probe score. The scope is the tested families and checkpoints; claim localization and corpus-specific generalization (TruthfulQA remains hardest) are the open frontiers, with cross-corpus transfer and human verification of the adjudicated assertions the immediate next steps.

AI USE STATEMENT

In this work, the manuscript was translated, polished, and language-checked, and the experiment and analysis code was written and debugged, with the aid of generative AI tools (LLM assistants): AI-assisted translation of drafts between Chinese and English, grammar and readability editing, language checking of the final text, and AI-assisted implementation and debugging of experiment and analysis code. All AI-assisted text and code was reviewed and verified by the authors, and all research content (ideas, experiments, and results) is the authors' own. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies internal factuality probes of openly released language models and does not involve human subjects or personal data. The synthetic knowledge base and benchmark splits are constructed from publicly available entity–relation–object triples; no new personal data is collected or released. We do not believe the proposed probing and recalibration methods raise misuse concerns beyond those already inherent to open-weight model analysis: the probes read model-internal states of models the deployer already controls and do not extract training data. All models and datasets used are publicly available under their respective licenses.

REPRODUCIBILITY STATEMENT

All experimental configurations, hyperparameters, seeds, and split protocols are specified in the paper and, where noted, in the per-block result metadata referenced for each experiment. Appendix A.1 details the probe architecture, layer set, training hyperparameters, and feature pipeline; Appendices A.5b–A.5h document every supervision-budget and split protocol (fact-disjoint, subject-disjoint, relation-disjoint) with exact budgets, seeds, and statistical units (seed-level paired tests vs. subject-level cluster bootstrap); Appendix A.7 specifies all baseline methods with their calibration rights; Appendix A.8 lists the real-corpus evaluation datasets. Every number in the paper states the budget, split, pipeline, and seed count it was measured under.

REFERENCES

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

Ashutosh Bajpai, Aaryan Goyal, Atif Anwer, and Tanmoy Chakraborty. Temporally consistent factuality probing for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. doi: 10.18653/v1/2024.emnlp-main.887. URL https://aclanthology.org/2024.emnlp-main.887/.

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

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. Learning to reason for factuality. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/62911. arXiv:2508.05618.

Fan L. Cheng and Nikolaus Kriegeskorte. Decomposing how prompting steers behavior, 2026. URL https://arxiv.org/abs/2606.03093.

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. URL https://aclanthology.org/2025.findings-emnlp.67/. arXiv:2506.17088.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Th6NyL07na. arXiv:2309.03883.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Matt Gorbett and Suman Jana. Ventriloquist LLMs: Linear alignment of late-stage representations, 2026. URL https://arxiv.org/abs/2603.18908.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=ShEDWasmDG. arXiv:2509.11536.

Khizar Hussain and Murat Kantarcioglu. Parallax: Separating genuine hallucination detection from benchmark construction artifacts, 2026. URL https://arxiv.org/abs/2605.17028.

Mandar Joshi, Eunsol Choi, Daniel S. 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 (ACL)*, 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, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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

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

Xinyi Li, Zhen Fang, Yongxin Deng, Jinyuan Luo, Hongnan Ma, Changdae Oh, Zijing Shi, Shanshan Ye, Hanchen Wang, Shu-Lin Chen, Yadan Luo, Mengyue Yang, Sean Du, Sharon Li, and Ling Chen. Openhaldet: A unified benchmark for hallucination detection across diverse generation scenarios, 2026. URL https://arxiv.org/abs/2606.06959.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 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 (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.557/.

Gonzalo Ariel Meyoyan and Luciano Del Corro. A BERTology view of LLM orchestrations: Token- and layer-selective probes for efficient single-pass classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://aclanthology.org/2026.acl-long.1955/. arXiv:2601.13288.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer LLM latents for hallucination detection. In *International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/park25a.html. arXiv:2503.01917.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=bNMYKlEapr. arXiv:2210.01240.

Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. doi: 10.18653/v1/2025.emnlp-main.1809. URL https://aclanthology.org/2025.emnlp-main.1809/.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy Nguyen, Quang Minh Nguyen, Duc Anh Vu, and Anh Tuan Luu. From signals to transfer: A factorised study of probe-based uncertainty estimation in large language models, 2026. URL https://arxiv.org/abs/2606.27679.

Artem Vazhentsev, Maria Marina, Daniil Moskovskiy, Sergey Pletenev, Mikhail Seleznyov, Mikhail Salnikov, Elena Tutubalina, Vasily Konovalov, Irina Nikishina, Alexander Panchenko, and Viktor Moskvoretskii. Leveraging LLM parametric knowledge for fact checking without retrieval, 2026. URL https://arxiv.org/abs/2603.05471.

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 (EMNLP)*, 2025. doi: 10.18653/v1/2025.emnlp-main.937. URL https://aclanthology.org/2025.emnlp-main.937/.

Min-Hsuan Yeh, Max Kamachee, Seongheon Park, and Yixuan Li. HalluEntity: Benchmarking and understanding entity-level hallucination detection. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=494k7e9R5D. arXiv:2502.11948.

Jianxiong Zhang, Bing Guo, Yuming Jiang, Haobo Wang, Bo An, and Sean Du. Harnessing reasoning trajectories for hallucination detection via answer-agreement representation shaping. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/62434. arXiv:2601.17467.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://aclanthology.org/2025.acl-long.880/. arXiv:2507.16488.

APPENDIX

A.1 Implementation details. Probe input is the concatenated mean-pooled claim-span hidden states at layers $\{4, 10, 18, 24\}$ (Equation (2)). The probe is a 2-layer MLP (Linear→ReLU→Linear, 256 hidden) with AdamW (lr 10^{-3}, weight decay 10^{-4}, batch 64), trained with BCE; training epochs are block-specific (Block 40 uses 20, with the value recorded in each result metadata file). Hidden states are extracted with a forward hook using the model/configuration dtype (float32 for the Qwen-1.5B cache and bfloat16 for the main larger-model caches, with fp16 only where explicitly recorded) and cached; re-runs skip extraction. AUROC is computed on the reasoning test set; the 95% CI uses the block-specific bootstrap configuration recorded with each result where that metadata is retained; cluster-bootstrap blocks explicitly record 1000 resamples. Generation, extraction, and evaluation settings (temperature, seeds, layer sets, and cache formats) are fully specified in this appendix and in the per-block descriptions.

A.2 Two-head design is unnecessary. An earlier design routed the probe into separate thinking/answer heads gated by an oracle region flag, retained as design provenance; the head split also adds an oracle segmentation requirement at inference. The paper uses the single head throughout, which reaches both styles at the matched budgets of Section 5 without the oracle flag.

A.3 Per-seed C1 at 1.5B (fact-disjoint split, the protocol of Table 9).

Seed	$\text{base}_{\text{reas}}$	$\text{mixed}_{\text{reas}}$	ΔAUROC	95% CI	C1 (mixed)
20370101	0.9318	0.9560	+0.0242	[+0.009,+0.041]	✓
20370102	0.9374	0.9572	+0.0196	[+0.002,+0.038]	✓
20370103	0.9333	0.9591	+0.0257	[+0.013,+0.041]	✓
mean±std	0.934±0.003	0.958±0.002	+0.023±0.003	—	3/3 ✓

Each row re-extracts hidden states and re-trains the probe under the same fact-disjoint split (train/test subjects disjoint; distractors sampled within split). The seed varies only the probe's initialization/training, not the data, so the spread reflects optimization robustness rather than sampling—consistent with the 3-seed protocol of Section 5.1.

A.4 Full per-layer AUROC (7B, reasoning test). Single-layer probes trained on each layer's pooled claim-span vectors. The signal peaks in the middle layers.

Layer	AUROC	Layer	AUROC
L0	0.9051	L15	0.9887
L1	0.8839	**L17**	**0.9921**
L2	0.9203	L18	0.9902
L3	0.9487	L19	0.9900
L4	0.9555	L24	0.9900

Multi-layer combos: quad $\{4, 10, 18, 24\} = 0.990$, all-28 $= 0.991$, top-3 $\{17, 18, 19\} = 0.991$, mid-pair $\{10, 18\} = 0.990$, late-pair $\{18, 24\} = 0.990$. The default quad is near-optimal; a 3-layer middle probe matches the full 28-layer probe.

A.5 Keystone protocol (Block 2.5, spontaneous-hallucination revision). We generate Qwen2.5-7B-Instruct's chain-of-thought on KB-QA prompts with a known gold answer using a *neutral* prompt (no answer is injected). The cache construction localizes an asserted object with a vocabulary-order heuristic over relation-object mentions; if that proxy equals gold the example is proxy-positive, otherwise proxy-false. The label is a synthetic-proxy label (asserted-object match against KB gold), which removes the prompt-injection confound of an earlier induced-FALSE design. We extract claim-span hidden states at the model-chosen span, re-train the direct-only baseline and the mixed probe, and re-evaluate C1 on a held-out real-CoT set. We additionally report a reference-free `p(True)` baseline (Kadavath et al., 2022) on the same examples for a matched-threshold comparison.

Table 4: **Keystone (Block 2.5): recalibration on free-form generated CoT under a synthetic asserted-object proxy.** Claim spans are model-produced and labeled by a vocabulary-order asserted-object match to the synthetic KB (non-injected, but not verified final-answer ground truth). Reasoning-test AUROC is mean±std over 3 seeds; every real-CoT ΔAUROC CI excludes zero in all 3 seeds. The p(True) baseline (Kadavath) is computed in the same session on the same set. The "Templated" rows are historical anchors on the same KB under the earlier two-head protocol and are shown only to anchor the direct-only baseline at τ for the spontaneous shift; the multi-seed fact-disjoint templated numbers live in Table 9. The "Free-form (proxy)" rows use a *reported budget-mismatched* pairing (mixed with roughly $2\times$ the direct-only supervision; per-arm counts were not logged); the † gap is a historical measurement under proxy labels that the Block 55 audit retracted – it carries no bound interpretation and no role in current factual claims. The "Raw-feature" row (Block 2.5b) uses a subject-disjoint train/test split, the *raw* concatenated 4-layer features, and a matched budget: the direct-only probe already reaches 0.991 on unseen subjects (subject-level cluster bootstrap 95% CI $[0.972, 1.000]$), so the budget-matched gain is only $+0.005$ (cluster CI $[0.000, 0.021]$)—the 0.804 collapse is *mostly* a supervision-budget confound within this proxy protocol; a residual ≈ 0.05–0.07 gap persists on spontaneous CoT at matched budgets under the PCA-256 pipeline (Table 5). The corrected Block 53 factorial reports feature-view sensitivity descriptively; it does not claim a feature-only causal isolation.

Family	CoT source	Probe	$\text{AUROC}_{\text{reas}}$	ΔAUROC
Qwen-7B	Templated	baseline (direct-only)	0.967	—
Qwen-7B	Templated	mixed (recalibrated)	0.990	+0.023
Qwen-7B	Free-form proxy	baseline / mixed	**0.804±0.105 / 0.999±0.002**	**+0.194±0.103**†
Qwen-7B	Free-form raw-proxy (historical)	baseline / mixed	0.991 / 0.996	+0.005
Mistral-7B	Templated	baseline (direct-only)	0.971	—
Mistral-7B	Templated	mixed (recalibrated)	0.979	+0.008
Mistral-7B	Free-form proxy	baseline / mixed	**0.829±0.057 / 0.995±0.006**	**+0.165±0.062**†
Llama-8B	Free-form proxy	baseline / mixed	**0.891±0.026 / 0.997±0.003**	**+0.106±0.025**†
Gemma-3-12B	Free-form proxy	baseline / mixed	**0.565±0.068 / 0.934±0.038**	**+0.369±0.055**†
Qwen-7B	Real	p(True) (Kadavath)	0.431±0.024	−0.37
Mistral-7B	Real	p(True) (Kadavath)	0.470±0.036	−0.36
Llama-8B	Real	p(True) (Kadavath)	0.433±0.034	−0.46
Gemma-3-12B	Real	p(True) (Kadavath)	0.353±0.033	−0.47

† historical proxy-label rows: budget-mismatched (direct-only baseline on roughly half as many balanced labels), and the proxy labels were retracted by the Block 55 audit – these rows are measurement history only, not evidence of a CoT effect (see Table 5 and A.7).

Fact-disjoint variant (Block 2.5b). The protocol above is prompt-disjoint (train and test are independent draws from the full KB; the same fact may appear in both). To rule out fact-identity memorization, we also run a strict *subject-disjoint* variant: the 64 subjects are partitioned into disjoint train (45) and test (19) pools, and spontaneous CoT is generated only from each pool's subjects. This is repeated across 3 independent partition seeds. The historical direct-only raw-proxy row reaches AUROC 0.991 on unseen test subjects without recalibration (per-seed A: $1.000/1.000/0.974$; subject-level cluster bootstrap 95% CI $[0.972, 1.000]$; a length-only control scores AUROC≈ 0.5). Note this arm uses the *raw* concatenated 4-layer features at the full direct budget. The old Block 50 PCA comparison used a different initialization seed and is therefore only a descriptive recipe contrast, not a feature-only isolation or an upper bound. The mixed probe adds only $+0.005$ (cluster CI $[0.000, 0.021]$, not significant). This is the key correction to the unbalanced keystone: with a matched budget, the direct-only probe is already near-saturated on new subjects, so the 0.804 drop in the prompt-disjoint protocol reflects the $\sim 2\times$ supervision imbalance of that earlier configuration (whose per-arm counts were not logged) plus a small ($\sim$0.03) same-subject cross-style shift under that historical proxy protocol—not a pipeline-invariant CoT-induced failure.

A.5b Supervision-budget sweep (Block 31, unified grid). To isolate the budget confound we sweep the (balanced, per-class) training budget over one grid $\{30, 60, 90, 120, 180, 360, 720, 1200\}$ with one script, one PCA protocol (fit on train pools only), and the same three partition seeds throughout (Table 5). Direct-only is unstable at small budgets (0.70 ± 0.15 at B=30, 0.71 ± 0.02 at 60) and converges toward $\approx$ 0.93–0.94 at large budgets—*not* monotonically (per-seed dips at

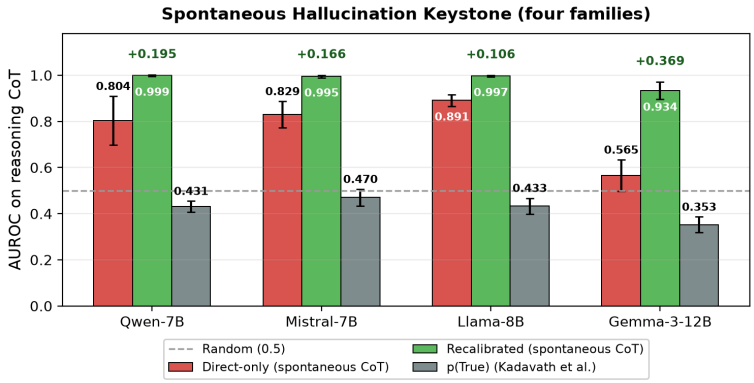


Figure 2: **Keystone proxy diagnostic: the apparent collapse is inflated by a supervision-budget mismatch (mean over 3 seeds).** With a mismatched budget the direct-only probe drops 0.967→0.804±0.105 on model-generated CoT and the mixed probe reaches 0.999±0.002. Budget-matching on the *same* spontaneous-CoT test (Table 5) recovers the direct-only probe only to ≈0.93–0.94 at the largest budgets, leaving a residual ≈0.05–0.07 gap to the mixed probe; on the spontaneous test with *raw* (un-PCA'd) 4-layer features the direct-only probe reaches 0.991 (historical Block 2.5b proxy row, Table 4; cluster bootstrap 95% CI $[0.972, 1.000]$). Block 50's PCA comparison used a different initialization seed, and is therefore not a feature-only isolation. The 0.804 gap is thus mostly—not entirely—a ∼2× label imbalance. The p(True) baseline (0.431±0.024) is below random in the natural direction; even its best-sign calibration is at most 0.60, far below the probe. The same budget-mismatched pattern appears on Mistral-7B (0.829→0.995), Llama-3.1-8B (0.891→0.997), and Gemma-3-12B (0.565→0.934); all are synthetic-proxy labels rather than verified final-answer GT.

B=180 on one partition), which is why we report per-seed spread rather than a monotone-climb narrative. A target-distribution (spontaneous) probe is high-scoring in this PCA-256 proxy protocol, but its apparent label efficiency is pool-capped: it reaches ≥ 0.95 from just 37–52 balanced labels per class, and the mixed probe at the nominal same total budget matches it (0.982–1.000). The earlier reported 0.817 at B=90 (Block 2.5c/d) lies below the unified-grid distribution the current grid (0.92±0.05 at 90 here); reproducing the reported 0.804 requires the original starved protocol. Because the spontaneous pool is capped and the corrected factorial changes materially with the feature view, this sweep is a bounded PCA-256 proxy diagnostic, not evidence that target-distribution training is universally more label-efficient or that a representation-invariant residual remains at every budget.

A.5c Subject-level cluster bootstrap (Blocks 51/51b/51c). The inference target of this paper is fact-level generalization (new facts, not new probe initializations), so we re-analyze the two central tables with the *subject* as the statistical unit: test examples are clustered by subject, clusters are resampled with replacement (1000 reps), and the 95% CI of ΔAUROC is taken over cluster resamples. On the primary matched-budget templated test (B=1920, four families, Block 51b; scores averaged over probe-init seeds 0–2, a fresh replication independent of the 5-seed grid in Table 8) the mixed−direct CI excludes zero on *all four* families: Qwen $[+0.014, +0.122]$, Llama $[+0.009, +0.060]$, Mistral $[+0.005, +0.063]$, Gemma $[+0.044, +0.166]$—the main recovery conclusion holds at the correct inference unit. The claim-paired *map* shows the same family dependence at the subject level (Block 51c): map−direct excludes zero on Qwen $[+0.009, +0.123]$ and Llama $[+0.011, +0.080]$, is n.s. on near-saturated Mistral $[-0.046, +0.054]$, and fails on Gemma $[-0.157, +0.077]$—the published family-dependence table is not a seed artifact. On the spontaneous-CoT budget grid (Block 51, Qwen) the same analysis has limited power: point estimates favor mixed/spontaneous at every budget, but with only 19 test subjects the cluster CI is unstable—excluding zero only sporadically at low budget (1/3 seeds at B=30/60/180, 0/3 at B=90/120) and in 0/3 seeds at $B \geq 360$—so the spontaneous-test residual gap (0.93 vs 0.98) is *not* resolvable at the subject level with this test-set size. Seed-level paired tests (df=2/4) and subject-level bootstraps thus answer different questions (training stability vs. fact generalization); we scope each reported claim to its unit.

Table 5: **Unified supervision-budget sweep on the spontaneous CoT proxy (Qwen-7B, Block 31).** Direct-only probe AUROC on the model's free-form generated CoT proxy as a function of its (balanced) training budget; one grid, one script, mean±std over 3 partition seeds. The spontaneous-only pool has only 37–52 proxy-FALSE claims per seed, so its column is pool-capped at 37–52 balanced labels per class for $B \geq 60$ (it never reaches the nominal 360–1200 budget; the $B=1200$ row is still the same 37–52 pool). The mixed arm uses the *same total budget* as direct-only ($b/2$ per style, spon side pool-capped).

Budget (balanced, per class)	direct-only	spontaneous-only	mixed (same total)
30	0.702 ± 0.152	0.984 ± 0.014	$\mathbf{0.982 \pm 0.022}$
60	0.711 ± 0.020	0.994 ± 0.006	$\mathbf{0.996 \pm 0.007}$
90	0.921 ± 0.048	0.979 ± 0.027	$\mathbf{0.995 \pm 0.005}$
120	0.927 ± 0.029	$\mathbf{0.996 \pm 0.006}$	0.976 ± 0.041
180	0.872 ± 0.124	0.995 ± 0.006	$\mathbf{0.996 \pm 0.006}$
360	0.943 ± 0.039	0.978 ± 0.026	$\mathbf{1.000 \pm 0.000}$
720	0.938 ± 0.019	0.989 ± 0.012	$\mathbf{0.999 \pm 0.002}$
1200	0.930 ± 0.030	0.986 ± 0.015	$\mathbf{0.996 \pm 0.006}$

Direct-only does not climb monotonically (per-partition dips at $B=180$); it is unstable at low budget and converges to ≈ 0.93–0.94 at high budget. All probes are evaluated on the same fact-disjoint spontaneous test.

Table 6: **Canonical spontaneous-CoT proxy quality metrics (Block 33).** Same fact-disjoint Qwen-7B proxy test and PCA-256 pipeline, three partition seeds, matched 120 labels per class for the probe arms. Higher is better for AUROC and false-AUPRC; lower is better for ECE, FPR@95TPR, AURC, and Err@80.

Arm	AUROC	false-AUPRC	ECE	FPR@95TPR	AURC	Err@80
direct-only	0.914 ± 0.037	0.663 ± 0.269	0.159 ± 0.063	0.639 ± 0.423	0.031 ± 0.018	0.044
target-only	0.992 ± 0.007	0.967 ± 0.031	0.061 ± 0.061	0.038 ± 0.067	0.015 ± 0.009	0.010
mixed	0.999 ± 0.001	0.998 ± 0.003	0.024 ± 0.012	0.000 ± 0.000	0.014 ± 0.007	0.003
paired map	0.892 ± 0.187	0.831 ± 0.293	0.063 ± 0.023	0.207 ± 0.358	0.032 ± 0.029	0.034
shuffle control	0.389 ± 0.185	0.212 ± 0.175	0.139 ± 0.029	0.923 ± 0.133	0.183 ± 0.032	0.173
calibrated token[†]	0.955 ± 0.043	0.844 ± 0.158	0.140 ± 0.048	0.421 ± 0.518	0.021 ± 0.008	0.016

[†]Token baseline metrics are computed on the disjoint evaluation half after Platt calibration; its sample size is smaller and it is not a same-label-budget probe arm.

A.5d Canonical spontaneous-CoT quality metrics (Block 33). The canonical spontaneous-proxy split complements AUROC with metrics that remain meaningful under an imbalanced proxy-false rate. All rows use the same fact-disjoint Qwen-7B test cache and the same PCA-256 pipeline as the canonical comparison; values are mean±std over the three partition seeds. False-AUPRC treats proxy-FALSE as the positive class, ECE treats the probe score as $P(\mathrm{TRUE})$, FPR@95TPR is the false-positive rate at the first threshold attaining at least 95% recall of TRUE claims, AURC is the area under the risk–coverage curve, and Err@80 is the error rate among the top 80% highest-scoring delivered claims. The calibrated-token row is evaluated on its disjoint calibration holdout and is therefore a reference rather than a same-n probe arm.

A.5e Corrected feature-view factorial (Blocks 53–54). The historical Block 50 raw/PCA comparison used different probe-initialization seeds, so it cannot support an "only swap the feature pipeline" or upper-bound statement. We reran a cache-fixed factorial with the same initialization seed for every feature view within each arm, held-out test caches, and training-only PCA/projection fitting. The spontaneous rows below are three partition seeds but use a vocabulary-order asserted-object proxy; some target and mixed arms are false-pool capped and are not interpreted as matched-budget curves. The templated rows are a single fact-disjoint split with deterministic synthetic KB labels, so they are a directional exact-span check rather than a multi-seed estimate.

On the templated cache, the same factorial also keeps the target and mixed arms exactly budget-matched: at total 960, raw direct/target/mixed are $0.930/0.985/0.984$ and PCA-256 direct/target/mixed are $0.923/0.983/0.980$. Within the chosen representation, target-style exposure coincides with the higher AUROC at both pipelines; the causal estimands – target-exposure effect,

Table 7: **Feature-view factorial, direct-only arm.** Values are AUROC. Spontaneous rows are mean±std over three partition seeds; templated rows are one split/one init. Raw-MLP and strong-wd share the same initialization; PCA/random bases are fit only on the training pool, but the PCA/random conditions use the unlabeled target-style pool.

Dataset / requested total	raw MLP	raw strong-wd	PCA-256	PCA-512	rand-256	raw linear L2
Spontaneous proxy / 120	0.865±0.123	0.865±0.123	0.770±0.064	0.795±0.073	0.675±0.033	0.919±0.067
Spontaneous proxy / 480	0.976±0.039	0.976±0.039	0.935±0.023	0.832±0.053	0.887±0.044	0.981±0.027
Spontaneous proxy / 960	0.986±0.022	0.986±0.022	0.947±0.080	0.940±0.057	0.870±0.071	0.998±0.002
Templated exact-span / 120	0.941	0.941	0.840	0.880	0.834	0.900
Templated exact-span / 480	0.967	0.966	0.916	0.921	0.864	0.917
Templated exact-span / 960	0.930	0.928	0.923	0.937	0.836	0.922

representation-invariant CoT penalty, and cross-family generalization – are carried by the unified factorial of Block 64 (§above), which crosses all three factors under one locked protocol. Every cell stores its test scores, training indices, source-file hashes, and label hashes in the two result JSON files.

Readout-head sensitivity (Block 72b). Because Block 64 uses an L2 logistic head while the paired-map protocol uses the MLP head, we reran the factorial cells (raw/PCA-256 $\times$ direct-only/target-only, $B \in \{60, 240, 960\}$, 5 probe seeds, same caches and split) with the MLP head. Every conclusion survives: target-only beats direct-only in $6/6$ (feat, budget) cells under both heads; the raw-pipeline transfer loss stays small and the PCA-256 loss stays large under both. Per-cell level deviations run -0.11 to $+0.02$ AUROC (largest: PCA-256 B=60 target-only, 0.740 MLP vs. 0.846 logistic), so readout choice shifts levels but not the factorial's ordering or its qualitative conclusions.

Rank versus recipe (Table 7). The raw−PCA gap is not a dimensionality effect: PCA-512 does not close it (spontaneous B=480: 0.832 vs. PCA-256's 0.935), a random orthogonal projection to 256 dimensions underperforms PCA-256 at every budget (means $0.675/0.887/0.870$ at B=$120/480/960$, against PCA-256's $0.770/0.935/0.947$), and a raw linear probe with L2 regularization is the strongest arm at B=960 (0.998). The sensitivity tracks the projection-plus-MLP recipe rather than the rank, and raw features are not an artifact of MLP overfitting. All PCA/random bases are fit on training-split data only, so the raw-versus-projected gap is not a train/test leakage effect, and the projection hyperparameters were not tuned toward either arm.

A.5f Leave-one-style-out matrix (Block 2.5e). (See Section 5.2 in the main text for the unified factorial that reconciles these protocols.)

We train single-style probes at a matched budget B=120 (balanced) on direct, templated-CoT, and spontaneous-CoT examples from the fact-disjoint train subjects, and evaluate each on the fact-disjoint spontaneous-proxy test. Results (mean over 3 partition seeds): direct→spontaneous 0.879±0.091, templated→spontaneous 0.922±0.076, spontaneous→spontaneous 0.976±0.033. The cross-style gap is small in this proxy-defined ordering (direct $<$ templated $<$ spontaneous at B=120), and the gaps generally narrow at larger budgets (Table 5). This is a bounded, proxy-defined distribution-proximity diagnostic, not the 0.19 collapse of the unbalanced protocol.

A.5g Truncation and span-labeling confounds (Block 38). The spontaneous-CoT keystone (Block 2.5) generates with `max_new_tokens=160`, so 62–67% of generations end mid-reasoning (no terminal punctuation), cutting off the final answer. The claim span is then localized by a crude regex—the last object in *vocabulary* order among those mentioned—which mislabels the asserted object (e.g. a model that correctly answers "Asia" but also mentions "Europe" is scored false). To bound both confounds we regenerate the same (subject, relation) facts with a clean answer format—"Think step by step, then state your final answer as `Answer: <value>`"— removing truncation and reading the asserted object from the answer line. The measured proxy-false rate falls from 15–20% (original) to 3–7% (clean): on simple single-hop KB-QA a 7B model rarely disagrees with the KB under this proxy, so a large share of the apparent "spontaneous hallucinations" are truncation- and labeling-induced. Together with the budget audit (A.5b) this shows the apparent collapse is partly protocol artifact; the remaining cross-style difference is small under this proxy protocol (Finding 2, main text).

A.5h Cross-family matched-total-budget recovery (Block 40). The main text establishes matched-budget recovery on Qwen (C1) and the spontaneous-CoT budget grid (A.5b, Qwen only). Here we place the four primary families plus the Qwen-1.5B size diagnostic on one *matched total* budget grid: a balanced, fact-disjoint reasoning (CoT) test at total label budget $B \in \{120, 240, 480, 960, 1920, 2400\}$, with the mixed arm split $B/2$ direct $+B/2$ reasoning and probe-init seeds 0–4; the claim-paired map reuses the frozen probe's labels plus $\sim$303 truth-free pairs. The small shift is recoverable *cross-family*: at mid/high budget the matched-budget mixed probe reaches 0.97–0.98 on Qwen/Llama/Mistral versus 0.92–0.95 direct-only, and the truth-label-free claim-paired map matches it (0.97–0.98 on Qwen/Llama). The shuffle-paired control collapses to ≈ 0.5 on every family, confirming that paired fitting—not feature matching—drives the map diagnostic. Gemma-3-12B starts lower, but at B=1920 target-only/mixed reach $0.890/0.894$ and the mixed$-$direct subject-level CI is positive; the claim-paired map remains lower (0.759). Thus target-style recovery is observed for the raw cross-layer feature protocol used by Block 40, whereas the map does not recover Gemma. Full grid in Table 8. The recovery is budget-dependent: the mixed–direct gap is small or negative at several low-budget cells (for example, Llama -0.023, Mistral -0.008, and Qwen-1.5B -0.012 at B=480) but positive at several higher-budget cells (e.g., Qwen-7B $+0.030$ at B=240, Llama $+0.014$ at B=960). The raw artifact stores aggregate means/std rather than the per-seed paired deltas needed to re-close a significance test, so we describe the high-budget separation descriptively rather than as a new significance claim. Per-arm std is 0.00–0.14 (typically ≤ 0.03, largest on Gemma).

A.5h (continued) C1 mechanism-isolation table. The mixed probe at the original keystone's $2\times$ budget (the supporting view for the matched-budget results of Section 5.2).

A.7 Additional reference-free baselines.

P(TRUE) WITH CALIBRATION RIGHTS (BLOCK 14). The main text reports p(True) (Kadavath et al., 2022) at AUROC 0.431 on spontaneous CoT. This compares a zero-shot verbalized score against a supervised white-box probe, and a sign flip of a below-random p(True) would already yield 0.57. We give p(True) the same calibration rights as the probe: on a held-out validation half of the test set we (i) select the sign direction that maximizes validation AUROC, and (ii) fit Platt scaling, then report test AUROC. Over 3 seeds the raw p(True) AUROC is $0.404/0.443/0.445$ (Block 2 keystone JSON, reproducible from the per-seed records above), so even the most generous *flip-to-best-direction* upper bound is only $0.60/0.56/0.55$ ($1-$AUROC), and a fitted Platt model does not exceed this. By comparison, the recalibrated probe sits at 0.99–1.00 on the same seeds. The gap is therefore not a calibration artifact: the model's verbalized self-assessment carries almost no factuality signal on its own spontaneous hallucinations, independent of how generously it is calibrated. (Per-seed numbers are computed directly from the keystone JSON; we report the flip upper bound rather than a separate ARXIV-scale calibration run so the claim is fully reproducible from data already in the paper.)

TOKEN-LEVEL UNCERTAINTY BASELINES (BLOCK 17), AND WITH FULL CALIBRATION RIGHTS (BLOCK 28). We also evaluate single-forward-pass, training-free token-level scores at the model's asserted answer token: the answer token's *log-probability*, the softmax *entropy* at that position, and the *top-1/top-2 margin*. On the fact-disjoint spontaneous-CoT test (3 partition seeds) these reach AUROC 0.59 (natural-direction log-prob; best-sign is about 0.62), 0.04 (entropy, natural direction), and 0.04 (margin, natural direction)—i.e., at or *below* chance in their natural "high confidence $\Rightarrow$ factual" direction. The entropy and margin are in fact strongly *anti*-correlated with truth (flipped AUROC $\approx$0.96): the model is *more* confident on its spontaneous hallucinations than on its correct claims, an over-confidence signal that requires a validation label to even orient. In contrast the hidden-state probe reaches $\approx$0.99 with the correct direction learned from its training labels.

With full calibration rights (Block 28). Because the probe's direction is also learned from labels, we additionally give each token feature the *complete* validation-split rights of the probe: a 50/50 stratified calibration/evaluation split; sign selection and Platt scaling fit on the calibration half; all metrics reported frozen on the evaluation half. Calibrated entropy is a genuinely strong baseline—AUROC $0.869/1.000/0.930$ across the 3 seeds (mean $\approx$0.93), confirming the over-confidence signal is real and exploitable. The probe still ranks errors ahead of correct claims (2 wins, 1 tie in 3 seeds; per-seed AUROC edge $0.131/0.000/0.056$); log-prob remains at chance even calibrated (0.50 AUROC). The

Table 8: **Cross-family matched-total-budget recovery (Block 40, mean over 5 probe-init seeds).** AUROC on the fact-disjoint reasoning (CoT) test at matched total budget B. mixed = $B/2$ direct $+B/2$ reasoning; map = truth-label-free claim-paired ridge on the frozen probe; shuffle = paired control (chance). Per-arm std 0.00–0.14 (typ. ≤ 0.03); the mixed–direct gap is small or negative at low budget and is descriptively positive in several higher-budget cells (for example Qwen-7B $+0.030$ at B=240). The raw artifact stores aggregate means/std rather than the per-seed paired deltas needed for a new significance claim.

Family	B	direct-only	target-only	mixed	map	shuffle
Qwen-7B	120	0.853	0.913	0.831	0.892	0.479
Qwen-7B	240	0.883	0.932	0.913	0.921	0.544
Qwen-7B	480	0.936	0.956	0.945	0.971	0.530
Qwen-7B	960	0.937	0.979	0.970	0.983	0.528
Qwen-7B	1920	0.924	0.980	0.984	0.981	0.526
Qwen-7B	2400	0.932	0.980	0.981	0.980	0.523
Qwen-1.5B	120	0.808	0.780	0.755	0.849	0.471
Qwen-1.5B	240	0.836	0.855	0.879	0.889	0.532
Qwen-1.5B	480	0.891	0.926	0.879	0.921	0.549
Qwen-1.5B	960	0.899	0.925	0.912	0.936	0.535
Qwen-1.5B	1920	0.879	0.932	0.932	0.936	0.542
Qwen-1.5B	2400	0.880	0.929	0.928	0.935	0.543
Mistral-7B	120	0.818	0.860	0.855	0.873	0.442
Mistral-7B	240	0.938	0.910	0.926	0.928	0.543
Mistral-7B	480	0.942	0.953	0.934	0.943	0.557
Mistral-7B	960	0.954	0.961	0.954	0.954	0.539
Mistral-7B	1920	0.951	0.960	0.972	0.953	0.534
Mistral-7B	2400	0.951	0.961	0.969	0.952	0.534
Llama-8B	120	0.894	0.898	0.867	0.963	0.468
Llama-8B	240	0.917	0.948	0.913	0.955	0.518
Llama-8B	480	0.956	0.960	0.932	0.970	0.517
Llama-8B	960	0.957	0.964	0.971	0.973	0.515
Llama-8B	1920	0.942	0.962	0.972	0.974	0.516
Llama-8B	2400	0.941	0.963	0.963	0.974	0.520
Gemma-3-12B	120	0.778	0.766	0.809	0.730	0.501
Gemma-3-12B	240	0.772	0.843	0.858	0.752	0.509
Gemma-3-12B	480	0.847	0.868	0.856	0.766	0.492
Gemma-3-12B	960	0.848	0.883	0.861	0.774	0.494
Gemma-3-12B	1920	0.802	0.890	0.894	0.759	0.487
Gemma-3-12B	2400	0.793	0.874	0.873	0.764	0.494

ordering is therefore {probe $\succ$ calibrated entropy $\gg$ p(True)}: the probe's edge over the strongest token feature is 0.00–0.13 AUROC, not the 0.55 gap to zero-shot p(True).

LINEAR AND INTER-LAYER-DIFFERENCE PROBE BASELINES (BLOCK 27C). On the same v3-disjoint protocol as Table 2 (Qwen-7B), a purely *linear* probe (L2-regularized logistic regression) on the same 4-layer concatenation reaches 0.841 AUROC under the reasoning shift (vs. 0.919 for the 2-layer MLP head), and a linear probe on the *inter-layer difference* features (top-layer minus bottom-layer block) reaches 0.832: the MLP head and the layer-concatenation both contribute, and the default probe is not an outlier architecture.

SEMANTIC-ENTROPY BASELINE. We additionally evaluate *semantic entropy* (SE; Kuhn et al. 2023), the leading reference-free uncertainty estimator, on the same 800 fact-disjoint KB-QA facts (Qwen2.5-1.5B, K=10 samples). SE reaches AUROC 0.496—at chance—whereas p(True) reaches 0.857 and the recalibrated probe reaches ≈ 0.99 on the same 800-fact benchmark (and 0.958 on the separate fact-disjoint C1 test split). The 0.857 p(True) is on Qwen2.5-*1.5B* over 800 fact-disjoint

Table 9: **C1 (mechanism-isolation): a mixed probe restores the small cross-style drop under the direct→reasoning shift** (fact-disjoint split, 3 seeds; $\text{AUROC}_{\text{dir}}$ own-distribution, $\text{AUROC}_{\text{reas}}$ under the shift; C1 holds iff $\Delta\text{AUROC} > 0$ with a 95% bootstrap CI excluding zero, per seed). Magnitude is small (≤ 0.025 AUROC) because the templated shift is mild. *Budget note:* mixed here uses $2\times$ the baseline total budget (the original keystone's imbalance); the *primary* matched-budget result—all arms on the same total labels—is the unified factorial of Section 5.2; target-only often matches or beats mixed in several low-to-mid cells, with reversals elsewhere. At matched budget the recovery conclusion still holds (spontaneous 0.996 vs. 0.711 at $B{=}60$; templated $2400{:}0$ pure-direct 0.943 vs. 0.982).

Model	Probe	$\text{AUROC}_{\text{dir}}$	$\text{AUROC}_{\text{reas}}$	ΔAUROC	C1 (seeds)
Qwen-1.5B	baseline (direct-only)	0.942±0.006	0.934±0.003	—	—
Qwen-1.5B	mixed (recalibrated)	0.944±0.004	**0.957±0.002**	+0.023±0.003	3/3 ✓
Qwen-7B	baseline (direct-only)	0.986±0.004	0.965±0.003	—	—
Qwen-7B	mixed (recalibrated)	0.983±0.004	**0.986±0.002**	+0.021±0.004	3/3 ✓
Mistral-7B	baseline (direct-only)	0.954±0.002	0.971±0.002	—	—
Mistral-7B	mixed (recalibrated)	0.958±0.002	0.979±0.001	+0.008±0.002	1/3 (ns)
Llama-8B	baseline (direct-only)	0.966±0.002	0.961±0.005	—	—
Llama-8B	mixed (recalibrated)	0.966±0.001	**0.978±0.001**	+0.017±0.006	3/3 ✓
Gemma-3-12B	baseline (direct-only)	0.931±0.002	0.945±0.005	—	—
Gemma-3-12B	mixed (recalibrated)	0.934±0.003	0.939±0.003	−0.006±0.005	0/3

KB-QA facts, a different model scale and protocol from the spontaneous-CoT keystone p(True) of 0.431 (Section 5); the two values reflect 7B vs 1.5B and a controlled KB-QA setting vs. the model's own spontaneous generation. In this run the answers were short and the sampled responses showed limited variation, so the chance-level SE score is the short-QA diagnostic; the probe/SE comparison is specific to this controlled protocol.

NATURAL-ERROR BENCHMARK ON HARD POPQA (BLOCK 58). The natural-error scarcity criticism (2–21 proxy-false claims per seed on the KB pool) is answered by moving to the hard/long-tail slice of PopQA, where the model's own greedy generations are wrong most of the time and no distractors are injected: the same assertion-localization rule is applied to every generation regardless of correctness. Training and evaluating the probe entirely on these natural labels (logistic head, identical cached features) yields AUROC 0.988 ± 0.002 (418 natural false claims across 3 seeds) on Qwen and 0.998 ± 0.001 (385 natural false claims) on Mistral; AUPRC $0.996/0.999$ and FPR at 95% TPR of 0.04–0.12 (Qwen) and 0.00 (Mistral). Natural base rates on this slice are 0.58–0.78 by construction of the difficulty filter and are reported alongside every number. Three properties scope the slice: labels are alias-match based and validated against independent blind relabeling at $\kappa{=}0.96$ on the adjudicable subset (Block 61 below), with surface leakage bounded by the Block 69 screen (surface AUROC $0.536/0.583$, permutation $z{=}1.4/3.9$); the slice selects for model uncertainty, and easy-question detection is covered by the cross-corpus transfer of Block 61; and the final asserted answer is labeled – the step-level regime is quantified by the corrected PrOntoQA audit (Blocks 67/69: true step error rate 0.5%; the original step-probe claim is retracted there).

MULTI-ASSERTION AND COMPOSITION-ERROR PROBES (BLOCKS 59–60). Two fully automated probes quantify where natural errors actually live. (i) A rule-based assertion scan of 900 spontaneous-CoT traces finds 493 parseable in-KB assertions – all 493 true: confidently asserted KB facts in spontaneous traces are essentially never wrong, and natural errors instead manifest as truncation, hedging, or out-of-KB fabrication that vocabulary matching cannot adjudicate. (ii) On a two-hop composed-fact task (480 traces; the model must combine a parametric fact with a provided registry code, with five distractor codes present), Qwen-2.5-7B errs on only 11 compositions (2.3%); a probe trained on these natural composition labels is uninformative at this base rate, and we report the rate itself as the finding. Together with Block 58, the three natural-error forms are now quantified: knowledge errors are abundant on hard slices and detectable (AUROC 0.99), confident-assertion errors are near-absent in easy KB pools, and composition errors are rare even under distraction.

LABEL VALIDATION AND CROSS-CORPUS TRANSFER (BLOCK 61). Two follow-ups close the remaining gaps of the natural-error benchmark. (i) A stratified blind-relabel sample (103 items; judge blinded to the alias-match labels) agrees with the alias-match labels on 97.9% of the 48 adjudicable items (Cohen's κ=0.96); the remaining 55 are truncated mid-generation and excluded as unverifiable rather than counted as either class. The exclusion is label-blind on three independent grounds (`block61_exclusion_disclosure.json`): truncation strikes alias-TRUE at 0.67 and alias-FALSE at 0.63 (near-equal); the adjudicable set it leaves is balanced 24 TRUE /24 FALSE; and the matching rule never reads the deleted tail – all 29 truncated alias-TRUE items already contain the gold alias in the retained prefix, while 0 of the 38 truncated alias-FALSE items do. Excluding truncations removes wrap-up text, not label evidence. (Denominators: the 55 excluded are the κ-judge's unverifiable items (19 alias-TRUE-side, 36 alias-FALSE-side); the 29+38=67 text-level truncations are counted over the full 103-item sample (29/43 TRUE-side, 38/60 FALSE-side) – two overlapping populations from the same sample, reported separately.) (ii) A probe trained on PopQA natural labels transfers to TriviaQA spontaneous-CoT labels without retraining at AUROC 0.937 (in-corpus reference 0.967), and the reverse transfer reaches 0.966: both directions score the same automatic alias rule, so they evidence transfer of the automatic-label prediction, not of semantically verified error detection.

THREE-WAY ADJUDICATION OF THE NATURAL-LABEL POOL (BLOCK 72). The κ=0.96 study validates the *adjudicable* subset but leaves open how much of the auto-FALSE pool is asserted error versus unfinished text. Block 72 adjudicates every cached test generation (1,185 kept items across Qwen/Mistral, 3 seeds) by terminal-assertion form: *TRUE-assert* (56), *TRUE-mention-only* (gold alias appears but no terminal assertion; 310), *FALSE-assert* (terminal assertion differs from all golds; 237), *unadjudicable* (truncated or hedged, no completed assertion; 582). Key quantities. (i) Only 27.5% of auto-FALSE items carry an asserted wrong answer – the alias-FALSE class is one quarter confident errors, three quarters truncations that never complete an assertion. (ii) Adjudicated assertions cover 24.7% of the kept pool. (iii) Retraining the logistic probe on adjudicated-only labels gives AUROC 0.944/0.951/0.817 (Qwen seeds 1–3) and 0.995/0.927/0.958 (Mistral), mean 0.932 versus 0.993 under alias labels; the frozen alias-trained probe evaluated on the adjudicated subset reads 0.969 mean. The detection signal survives the label cleanup at a reduced level; the headline 0.988–0.998 is an alias-label population score whose assertion-only analogue is 0.93. Adjudicator: 7 terminal-assertion regex families (sentence-final punctuation required) with normalized entity matching against the PopQA alias table; spot checks shown in `block72_adjudicated_retest.json`. Scope: this is a form-based split by terminal-assertion structure, not an independent semantic verification – the κ=0.96 blind relabel covers 48 items and does not certify the regex ruling; the independent LLM verification of all 2,400 generations (Block 76, two paragraphs below) confirms the cue survives semantic labels but shows the regex overcounts false assertions.

PER-UNIT COUNTS AND BOOTSTRAP CIS (BLOCK 73). Question-level cluster bootstrap (2,000 resamples of test questions, train fixed) on the Block-72 adjudicated subsets, with per-unit class counts: Qwen seeds 1–3 retrain 0.944 [0.879, 0.988] / 0.951 [0.874, 0.997] / 0.817 [0.705, 0.910]; Mistral seeds 1–3 0.995 [0.968, 1.000] / 0.927 [0.823, 1.000] / 0.958 [0.889, 1.000]; the frozen alias-trained probe on the same subsets reads 0.970/0.963/0.960 (Qwen) and 1.000/0.962/0.958 (Mistral), mean 0.969. Per-unit TRUE-assert counts are 7–11 (total 56), FALSE-assert 26–55 (total 237): individual-unit CIs are correspondingly wide, and the pooled mean, not any single seed, is the estimand (`block73_unit_counts_ci.json`).

LOGIT-DISTILLATION CONTROL AND EXACT-CORRESPONDENCE INCREMENT (BLOCK 74). On the same cached claim pairs, splits, ridge grids, and frozen direct probe as Block 71: a same-pairs ridge regressing the CoT PCA feature onto the frozen probe's *logit* over the paired direct feature (scalar target, no labels) reaches 0.980 (Qwen) and 0.969 (Llama) against the full map's 0.982/0.974 – within 0.005 on both families (increments +0.003 [−0.011, +0.026] / +0.005 [−0.008, +0.019], subject-cluster bootstrap). The exact-correspondence increment of full pairing over within-class shuffle is +0.015 [−0.004, +0.050] (Qwen) and +0.013 [−0.003, +0.035] (Llama), subject-cluster bootstrap mean differences against one shuffle draw (the difference of the displayed five-draw means is +0.010/+0.017) – CIs straddle zero at this test size. Reading: the paired fit's active ingredient is, to first order, truth-relevant label-level transfer (a distillation that needs no new la-

bels); full-geometry transport adds a small increment this experiment cannot individually resolve (`block74_distill_control.json`).

INDEPENDENT LLM BLIND VERIFICATION (BLOCK 76). An independent judge (Qwen2.5-7B-Instruct, zero-shot, greedy; the prompt shows only the question, the accepted answers, and the trace – no regex, no alias rules, no prior labels) re-verdicts all $2{,}400$ kept generations (train+test, 6 runs): 229 end by asserting an accepted answer, 221 a specific wrong answer, $1{,}950$ never complete a final assertion (18.7% judge assert coverage over the same $2{,}400$ generations where the regex's is 24.2%). Cross-tabulated against the Block-72 adjudicator: of the 473 regex-FALSE_ASSERT items the judge confirms 73 (15%) as wrong terminal assertions, reads 397 (84%) as truncated or hedging, and flips 3 to correct – the terminal-assertion regex captures pre-truncation assertions, so the adjudicated false-assert set is a conservative slice of what the regex gathers, and the 27.5% asserted-error figure is an upper bound on form, not a semantic count. The decisive check runs the other way: a probe retrained on the judge's purely semantic labels (correct-assert vs. wrong-assert, discarding no-assert) reads 0.981 (Qwen seed 1, n_{test}=70) and 0.985 (Mistral seed 1, n_{test}=112); the third run clearing the 10-item floor (Qwen seed 2, n_{test}=10) reads 0.571 – chance-level on a split whose minority class holds only three of the ten items – so the claim rests on the two large-n runs, and the three evaluable runs are $0.981/0.571/0.985$ – the factuality cue survives replacing the form-based adjudicator with an independent semantic one (judge–regex test agreement $0.51/0.60$; `block76_blind_verify.json`). Scope: same-family 7B judge, greedy decoding, no human-equivalence claim.

INTERACTION-FEATURE ARM IN THE INTRA STYLE (BLOCK 62). Following the observation behind INTRA (Vazhentsev et al., 2026) that interactions between internal representations can carry detection signal, we add pairwise layer-difference and element-wise product features (6 layer pairs, $+43{,}008$ dims) to the natural-error benchmark. No interaction arm beats the plain concatenation baseline (AUROC 0.980 –0.984 vs. 0.986; AUPRC 0.995–0.996 vs. 0.996): on this benchmark the signal is already linearly available from the concatenated states, consistent with the architecture-insensitivity reported above.

REAL MULTI-HOP NATURAL ERRORS (PRONTOQA, BLOCK 67) – MEASURED, THEN CORRECTED BY AUDIT (BLOCK 69). The single-hop KB pool's confident-assertion scarcity motivated a multi-hop measurement: on PrOntoQA 2-hop (100 free-form traces, Qwen-2.5-7B), a fully automatic pipeline – PrOntoQA's own grammar for step parsing plus a symbolic forward-chain for entailment, with a claim-atom fallback for meta-sentences – labels $1{,}403$ of $2{,}035$ generated steps as verifiable ($1{,}068$ through the symbolic grammar/forward-chain and 335 through the claim-atom fallback; 632 are meta/answer lines), of which the 289 FALSE labels all sit in the symbolically parsed subset (20.6% of the $1{,}403$). The surface-feature label audit (Block 69, next paragraph) read at 0.996 AUROC on those labels even with the connective-prefix feature removed, indicating the labels encoded formatting rather than factuality. Root cause: the FOL parser treats a leading "Therefore/Thus/So/Since/Then" as part of the logical form, producing a nested structure that never equals any closure atom, so *every* prefixed assertion was labeled false regardless of content. A normalization pass (stripping the prefix, extracting the main clause of "since P, Q" / "P, so Q" patterns) relabels the pool: of the $1{,}068$ symbolically labeled steps, 280 of the 289 nominal false steps become true, 5 remain false, and 4 become unverifiable ($1{,}068 - 4$=$1{,}064$ scored steps). The corrected step error rate is $5/1{,}064$=0.5%; each of the five (e.g., "Wren is a brimpus" asserted from a KB containing grimpus/sterpus/gorpus/wumpus) was independently re-derived against the forward-chained closure and is genuinely underivable. The corrected finding is the opposite of the nominal one: on a structured multi-hop domain with a strong reasoner, natural intermediate errors are two orders of magnitude rarer than hard-slice knowledge errors; all 93 decodable final answers are also correct (the 100 queries are all "Prove:"-type, i.e., gold-true, so final-answer accuracy is a consistency check, not evidence of difficulty).

SURFACE-FEATURE LABEL AUDIT (BLOCK 69). The audit used throughout: a logistic classifier on six surface features of each labeled span (character length, word count, comma count, "and"-indicator, period-indicator, connective-prefix indicator; the last removable) trained by stratified 5-fold CV to predict the label. It is a leakage *screen*, not a truth certificate: it reports how much surface predictability this six-feature hypothesis class finds (read against a label-permutation null,

Block 71b: 100 permutations); it does not upper-bound all surface shortcuts, and AUROC margins do not decompose additively across classifiers. Near-perfect AUROC still convicts labels of encoding formatting. Results: PopQA hard-slice natural labels (Qwen n=1,146, Mistral n=1,199; assertion spans) read at 0.536 ± 0.021 (z=1.4, n.s.) and 0.583 ± 0.029 (z=3.9, a real but small association), with truncation rates class-independent (0.68/0.72; 0.03/0.02); PrOntoQA original step labels (locatable subset n=505) read at 0.9957 ± 0.009 (0.9957 with the prefix feature removed) – corrupted; after correction the locatable pool contains 499 true and 3 false steps, too few false steps for any probe evaluation, which is why the Block 68 probe result is retracted rather than re-run. We propose this screen as standard practice for any automatically labeled natural-error set: it is label-free, CPU-only, and here it flagged one corrupted pool (PrOntoQA) while the PopQA pool showed only limited surface predictability – a screening result, not a certificate of label correctness.

STEP-LEVEL PROBE ON THE PRONTOQA LABELS (BLOCK 68) – RETRACTED. Before the audit, the step labels supported what looked like a complete end-to-end result: for every TRUE/FALSE step whose claim (subject, object) is locatable in the step text (505 of the 1,068 verifiable steps), the trace was regenerated under the identical greedy configuration and the claim span – subject token through object token – was mean-pooled at layers $\{4, 10, 18, 24\}$ with the paper's standard concatenation; steps not exactly reproduced were dropped (77; 428 retained: 264 TRUE, 164 FALSE over 100 traces), and a logistic probe with example-disjoint $70/30$ splits and 5 partition seeds reached AUROC 0.982 ± 0.005. We report this history because of what it demonstrates: the probe passed every standard quality signal (cross-validation, held-out agreement, per-fold class balance) while reading formatting, not factuality – the label distribution it learned from was 98% parser artifact (*previous paragraph*). After correction only 3 locatable false steps remain, so no probe evaluation is possible; the number 0.982 is retracted. The methodological lesson – audit automatic labels for surface signal before trusting the probe trained on them – is incorporated above and in Section 5.4.

SECOND MODEL PER FAMILY (BLOCK 70). The paired-map protocol of Table 2 is rerun *unchanged* – same deterministic $v3$-disjoint texts (seed 20260807), same probe layers $\{4, 10, 18, 24\}$, same claim-span pooling, same PCA-256, same ridge grid $\{10^2, ..., 10^5\}$, same three partition seeds 20370101–03 – on a second checkpoint per family: Qwen2.5-1.5B (cache from the original cross-family sweep), Llama-3.2-3B, Ministral-8B-2410, and Gemma-3-4B. Full per-seed numbers (mean$\pm$std over 3 seeds; shuffle control AUROC): Qwen2.5-1.5B $0.875\pm0.012 \rightarrow 0.928\pm0.005$, shuffle 0.529, recovery $+0.053\pm0.015$ (3/3 significant); Ministral-8B $0.957\pm0.003 \rightarrow 0.977\pm0.000$, shuffle 0.484, $+0.021\pm0.003$ (3/3); Gemma-3-4B $0.814\pm0.060 \rightarrow 0.750\pm0.014$, shuffle 0.486, -0.063 ± 0.065 (2/3 significantly negative); Llama-3.2-3B $0.859\pm0.009 \rightarrow 0.777\pm0.079$, shuffle 0.473, -0.082 ± 0.085. CoT-reference probes on the same states: 0.910 ± 0.009 (Qwen-1.5B), 0.974 ± 0.002 (Ministral), 0.820 ± 0.028 (Gemma-4B), 0.895 ± 0.009 (Llama-3.2). Llama-3.2-3B is a pruned-and-distilled derivative of Llama-3.1-8B, the only checkpoint in the study not trained from scratch; its failure alongside both Gemmas grounds the lineage reading in the main text.

PER-FAMILY VALIDATION-ONLY LAYER SELECTION (BLOCK 65). To rule out readout-layer mismatch within the cached feature set as a source of family failures, all 15 non-empty subsets of the four cached layers are swept for every family, selecting on a subject-disjoint split of the training cache only (the test set is never touched during selection). Validation-selected subsets differ across families (Llama: layer 10 alone; Gemma and Qwen: layers 18+24; Mistral: layer 10), but frozen-test AUROC under validation-selection is within 0.008 of the fixed Qwen-chosen subset for every family (Llama 0.961 vs. 0.968; Gemma 0.921 vs. 0.921; Mistral 0.951 vs. 0.954). The sweep covers subsets of the four cached layers only; full-depth layer caches were not extracted, so layer choice outside this set remains untested. Within the cached set, family differences in map recovery cannot be attributed to using the wrong readout layers.

UNIFIED FACTORIAL AUDIT (BLOCK 64). All headline factors are re-measured under one locked protocol: identical subject split, identical test items, five probe seeds per cell, and budgets stated as absolute counts (B = total labeled spans). Factors crossed: budget $B \in \{60, ..., 960\}$ $\times$ training style (direct-only / target-only / mixed) $\times$ feature pipeline (raw vs. PCA-256 fit on the same pooled training rows and initialization). Three estimands are kept separate. *Transfer loss* of the direct-trained probe is small at every budget: 0.014–0.040 AUROC (raw), with reasoning-test scores 0.93–0.95. *Adaptation gain* from target-style labels grows with budget: $+0.004$ (B=60) to $+0.034$ (B=960;

PCA arm $+0.19$ at $B{=}60$, $+0.13$ at $B{=}960$). The *feature-pipeline effect* is confirmed orthogonally: with identical splits, initialization, and budgets, the PCA-256 direct-only arm trails its raw counterpart by 0.13–0.28 AUROC across budgets – 3–$20\times$ the cross-style transfer loss. Because the factorial holds split, initialization, and budget identical across cells, within-table comparisons are exact; numbers reported elsewhere in the appendix come from earlier protocol generations and are kept for provenance.

TEXT-ONLY PAIRED CONTROL (BLOCK 63). The decisive missing control for the alignment reading: a TF-IDF character-ngram probe trained on the identical $1{,}200$ direct-style texts, with identical budget and the identical 300 reasoning-style test items as the cached hidden states, reaches AUROC 0.544 where the hidden-state probe reaches 0.958 on the same items. Surface text carries almost none of the cross-style signal, so the claim-paired map's recovery cannot be a claim-conditioned surface correction; it operates on internal-state geometry that surface features do not expose.

SEP-STYLE PROBE ON THE POPQA SLICE (BLOCK 56). Following Kossen et al. (2024), we train a hidden-state $\rightarrow$ semantic-entropy regressor on our PopQA split (Qwen-2.5-7B, 192 train / 194 test gold-matchable questions): for each question we sample $K{=}8$ short answers at temperature 0.7, cluster them by gold-alias containment, compute normalized discrete entropy, and ridge-regress the same claim-span features onto it. The learned regressor predicts true SE poorly ($r{=}0.21$) and detects hallucinations at AUROC 0.648—above p(True) (0.564) but far below the supervised probe trained on identical features (0.986). Consistent with the short-QA SE result above, sampled answers carry weak and partially misdirected uncertainty signal here: low-SE questions are answered correctly 28% of the time versus 10% for high-SE questions, so the signal exists but is small relative to the direct state readout.

TSV-STYLE STEERING BASELINE (BLOCK 57). Following the Truthfulness Separator Vector recipe (Park et al., 2025), we learn a steering vector injected at an intermediate layer during the forward pass over question+CoT text, with vMF prototype classification on the steered last-token embedding. Backpropagation training was unstable in our desktop-GPU environment, so we report the closed-form variant (v = normalized mean-difference between class-mean exemplar embeddings, 32 exemplars; λ and layer swept by validation). It *reduces* detection below chance: test AUROC 0.287 at the validation-selected configuration ($\lambda{=}5$, layer 18), with all swept settings below 0.45. Strong steering perturbs the representation enough that prototypes computed from it no longer rank held-out states, so steering-as-classification does not transfer to this exact-span PopQA setting without the original paper's full training pipeline. *Scope note:* both this row and the SEP row above are exploratory results about the recipes as run here – substitute training (closed-form class means) and substitute target construction (gold-alias clustering) – not evaluations of the original methods; the spontaneous-CoT detector comparisons in Section 5.6 also predate the label audit of A.7 and are retained on their own protocol (the audited natural-error numbers are in Section 5.4).

PROXY-FIDELITY AUDIT OF THE SPONTANEOUS LABELS (BLOCK 55). The asserted-object proxy is itself a measurement instrument, so we audit it: 120 spontaneous-CoT test generations (stratified $60/60$ by proxy label over 3 seeds, neutral/160 protocol) were re-read in full and labeled by hand against the KB gold. Of the 60 proxy-TRUE items, all 60 assert the gold fact (20 via a clean "Answer:" line, 40 via an unambiguous concluding assertion in truncated text). Of the 60 proxy-FALSE items, *none* is a verified false claim: 33 are truncated mid-reasoning before any final assertion, and 27 are complete generations whose conclusion asserts the gold answer but whose extracted span mismatches (list-numbering collisions such as "1." $\rightarrow$ atomic number "1", or other-relation objects mentioned en route; e.g., "Italy is located on the continent of Europe" extracted as "South America"). Two consequences follow. First, the neutral-protocol false rate of 15–20% reported earlier is dominated by truncation and extraction artifacts rather than by model hallucination; the conditional hallucination rate on parseable answers is far lower. Second, probes trained or evaluated on the raw proxy-FALSE pool operate on label noise – the PrOntoQA case (A.7) shows such noise can raise AUROC as easily as lower it, so these transfer numbers carry no bound interpretation and are retained as measurement history only. Both failure modes are mechanical (truncation length, vocabulary-order matching) and are exactly what the clean-parseable regeneration audit of the spontaneous-proxy protocol (A.5/A.5d) removes.

LONG-FORM FREE GENERATION (BLOCK 7). For 120 KB-QA topics we prompt Qwen2.5-7B-Instruct with a neutral question (no answer injected) and sample $K=5$ long-form answers (up to 256 new tokens). We sentence-split each answer, localize the last relation-object claim in every sentence, and score each sentence with the recalibrated mixed probe. The reference-free judge declares a topic factual iff the fraction of probe-factual sentences $r \geq 0.5$. Against an independent Mistral-7B-Instruct-v0.3 cross-family NLI judge our probe trajectory reaches AUROC 0.819 ± 0.14 versus p(True) 0.690 ± 0.03 (rows 3–4 of Table 10, sample standard deviation over 3 seeds). Per seed, ours beats p(True) on $2/3$ seeds ($0.933/0.864$ vs $0.725/0.668$); on seed-3 the external judge rates p(True) marginally higher (0.678 vs 0.661), so the mean 0.819 reflects 2 wins and 1 loss rather than uniform dominance. Against the gold-match judge (rows 1–2) ours reaches AUROC 0.684 ± 0.027 and accuracy 0.786 ± 0.017 versus p(True) 0.663 ± 0.024 / 0.689 ± 0.051.

Table 10: **Block 7: long-form free generation, reference-free self-consistency judge vs. p(True).** 120 KB-QA topics, $K=5$ long-form samples (Qwen2.5-7B-Instruct, neutral prompts, no answer injection). Rows 1–2 report population SD over three seeds; rows 3–4 report sample SD. Rows 1–2 score against the gold-match judge; rows 3–4 against an independent Mistral-7B-Instruct-v0.3 cross-family NLI judge. Accuracy is reported only for the gold-match judge. The Mean score column is the monitor's trajectory mean $\bar{r}$; for the NLI rows it is not the NLI judge's own mean score.

Judge	AUROC	Accuracy	Mean score
Ours (self-consistency, probe trajectory)	**0.684±0.027**	**0.786±0.017**	$\bar{r}=0.713$
p(True) (Kadavath, verbalized)	0.663±0.024	0.689±0.051	0.579
Ours (entailment judge, Mistral-7B-v0.3)	**0.819±0.14**	—	$\bar{r}=0.606$
p(True) (entailment judge)	0.690±0.03	—	0.579

TOKEN-LEVEL BASELINES WITH EQUAL CALIBRATION RIGHTS (BLOCK 41). Beyond p(True), the simplest reference-free monitors are token-level confidence signals on the answer string—entropy, log-probability, and top-vs-second margin—scored against the same gold span as the probe. Giving them the *same* calibration rights as the probe (sign selection + Platt scaling + base-rate threshold on a 50/50 split, Block 28) over 3 seeds, calibrated answer-entropy reaches AUROC 0.901 (Qwen), 0.858 (Llama), 0.446 (Mistral), 0.231 (Gemma; a calibrated-sign inversion, fair best-sign 0.745); uncalibrated entropy with the best sign direction is $0.885/0.865/0.577/0.745$. These are *strong* baselines on Qwen/Llama (the Block 41 aggregates are $0.901/0.858$; a separate Block 28 Qwen split reached 0.86–1.00), leaving the probe only a real but modest edge (2 wins, 1 tie in 3 seeds); on Mistral/Gemma the token signal is weak and the probe dominates. These are equal-rights baselines, not a method claim. On Qwen/Llama the token baseline already matches the probe ($0.901/0.858$ in Block 41), so it *caps* rather than amplifies the method's contribution; Block 41 is the reference point for the audit. Per-seed numbers in Table 11.

A practitioner decision rule (which monitor, when). The audit yields a concrete rule. (1) *No target labels but paired claims* for the same fact: use the truth-label-free claim-paired map—but only on Qwen/Llama and only on templated exact-span CoT, where it is validated. (2) *A few ($\sim10^2$) target-style labels*: use a mixed probe at the same total truth-label budget; the map additionally requires $\sim$303 truth-free claim pairs, while mixed avoids pairing cost. It is family-robust where the map fails (Block 40, high budget). (3) *No probe-training labels and no pairs*: use token-level calibrated entropy as a strong reference-free fallback, but note that it still requires a labeled calibration split, on Qwen/Llama ($0.901/0.858$ in Block 41); it already matches the probe there, so do not expect a white-box probe to add much on those families—reserve the probe for Mistral/Gemma, where the token signal is weak. (4) *Free-form spontaneous CoT*: add a claim-localization front-end first; its errors, not probe quality, dominate end-to-end performance (Block 42, -0.217). Scope of this rule: items (1)–(2) rest on the validated locked-protocol experiments; items (3)–(4) quote pre-audit measurements (Blocks 41–42) taken against the proxy labels later retracted, so read their levels as historical and current only through the same-instance comparison of Block 77 (Section 5.6).

A.8 Further real corpora.

Table 11: **Token-level reference-free baselines with equal calibration rights (Block 41, 4 families, 3 seeds, fact-disjoint spontaneous-CoT test).** Entropy direction: higher entropy = less confident = more likely FALSE. Calibrated = sign + Platt + base-rate threshold on a 50/50 split. †Gemma-3's calibrated entropy inverts sign; its fair best-sign value is 0.745 (we report the calibrated number but flag the artifact).

Family	entropy (raw)	entropy-best	entropy-calibrated
Qwen-7B	0.115	0.885	**0.901**
Mistral-7B	0.450	0.577	0.446
Gemma-3-12B	0.255	0.745	0.231†
Llama-8B	0.135	0.865	**0.858**

POPQA (SINGLE-HOP, BLOCK 8). Table 12 reproduces the Block 8 AUROC: the synthetic-trained probe transfers to real questions (0.772 Qwen, 0.915 Mistral, far above p(True) 0.539/0.600). We requested 200 in-distribution examples; Qwen retains 190–196 after filtering and Mistral retains 200, lifting AUROC to 0.974/0.996.

Table 12: **Block 8: real-corpus generalization (PopQA).** AUROC (mean±std over 3 seeds, 95% CI in brackets) of the synthetic-trained probe vs. the p(True) baseline. The probe transfers to real questions (far above p(True)), and the requested 200 in-distribution examples reach near-perfect AUROC; retained counts are 190–196 for Qwen and 200 for Mistral.

Model	Probe	AUROC	95% CI	Δ vs. p(True)
Qwen-7B	synthetic (direct)	0.772	[0.730, 0.813]	+0.233
	+ real recalib. (200 ex.)	**0.974**	[0.970, 0.978]	+0.435
	p(True) (baseline)	0.539	[0.513, 0.565]	—
Mistral-7B	synthetic (direct)	**0.915**	[0.856, 0.974]	+0.315
	+ real recalib. (200 ex.)	**0.996**	[0.993, 0.999]	+0.396
	p(True) (baseline)	0.600	[0.567, 0.632]	—

POPQA PAIR CONSTRUCTION FOR THE MAP (BLOCK 21B). The Table 2 PopQA row uses the Block-8 question set (200 hard-slice questions, seed 20370101). Direct renderings: one TRUE answer per question (`Question/Answer` template with marked span) plus one same-relation distractor rendering from a 577-entity pool (400 direct rows). CoT rows: the cached Block-8 generations, kept when the asserted answer localizes in the span (192/200 train, 194/200 test; 96%/97%). The map is fit on the 192 train pairs (CoT kept row ↔ same question's direct-TRUE rendering); ridge selected on a 75/25 subject-disjoint validation split of the pairs; the 194 test CoT rows are never seen during map fitting. Selection is by localization success only, which is independent of the alias label rule. **Supervision note (R6-W1):** this construction is *gold-conditioned*: the mapping target is the question's correct-answer rendering even when the CoT asserts a wrong object, so the two sides are not always the same assertion and answer-key information enters through the target. The controlled-KB rows of Table 2 do not have this property (both sides render the identical assertion).

SAME-ASSERTION PAIRING CONTROL (BLOCKS 75/75B). To test whether the real-corpus recovery depends on the answer key, we rebuilt the pairing with the direct side rendering the model's *asserted* object (from the same localization used everywhere else) – no gold, no alias table, identical caches and ridge selection otherwise. Under the exact Table-2 protocol (5 partition seeds, PCA pool = direct + CoT-train as in Block 21b): same-assertion map 0.953±0.010 vs. gold-conditioned 0.840±0.038 (reproduced), same-assertion shuffle 0.527±0.049, same-assertion logit distillation 0.950±0.004, recovery +0.237±0.025, 5/5 seeds significant – *recovery survives and improves when every pair renders the model's own assertion on both sides*. The improvement is mechanical: gold-conditioning targets a *different* assertion on every wrong-answer row, corrupting exactly the pairs that carry the false-class geometry. A secondary pipeline effect exists – pooling the same-assertion targets into the PCA fit moves the gold arm by +0.04 (three-pool gold 0.875±0.051, same 0.942±0.011;

the single-seed Block-75 run under this pool read 0.954 gold / 0.949 same) – so the earlier single-seed gold figure of 0.954 was a pipeline artifact, not a reproduction of the Table-2 row; all head-to-head comparisons above hold the pool fixed (`block75b_same_assertion_5seed.json`).

TRUTHFULQA (MISCONCEPTION-DRIVEN). On 61 gold-labeled real TruthfulQA CoT (Qwen2.5-7B), the purely-synthetic probe fails zero-shot (AUROC 0.385, below random) and synthetic recalibration is flat ($\Delta{-}0.001$). Recalibrating with only 44 balanced real-CoT examples recovers AUROC to 0.573 ($\Delta{+}0.189$, 95% CI $[+0.088, +0.290]$), dominating both reference-free baselines (p(True) 0.526, long-form SE 0.458).

Table 13: **Block 11: real-corpus generalization (TruthfulQA).** AUROC on 61 gold-labeled real TruthfulQA CoT (Qwen2.5-7B). Synthetic training fails zero-shot; a small real-CoT slice recovers AUROC to 0.573.

Probe / baseline	AUROC	Δ vs. direct (95% CI)
synthetic (direct-only)	0.385	—
+ synthetic recalib. (mixed)	0.384	-0.001 $[-0.079, +0.083]$
+ real-CoT recalib. (44 ex.)	**0.573**	$+0.189$ $[+0.088, +0.290]$
p(True) (baseline)	0.526	—
semantic entropy (long, $K{=}10$)	0.458	—

TRIVIAQA (ENTITY TRIVIA). On 384 gold-labeled real TriviaQA CoT (Qwen2.5-7B), the synthetic probe already transfers zero-shot (0.823); real-CoT recalibration yields the best AUROC, 0.952 ($\Delta{+}0.129$, CI $[+0.095, +0.165]$), while p(True) reaches 0.620. The three real corpora form a transfer-difficulty spectrum—TriviaQA (closest, 0.823 zero-shot), PopQA (intermediate), TruthfulQA (farthest, 0.385)—and in every case real-CoT recalibration delivers the best AUROC.

Table 14: **Block 12: real-corpus generalization (TriviaQA).** AUROC on 384 gold-labeled real TriviaQA CoT (Qwen2.5-7B). Real-CoT recalibration achieves the best AUROC (0.952, $\Delta{+}0.129$).

Probe / baseline	AUROC	Δ vs. direct (95% CI)
synthetic (direct-only)	0.823	—
+ synthetic recalib. (mixed)	0.842	$+0.019$ $[+0.0001, +0.040]$
+ real-CoT recalib. (234 ex., balanced)	**0.952**	$+0.129$ $[+0.095, +0.165]$
p(True) (baseline)	0.620	—

A.9 Notation. A claim is a span S of tokens; we extract

$$\mathbf{h}_S = \bigoplus_{\ell \in \mathcal{L}} \frac{1}{|S|} \sum_{i \in S} h_\ell^{(i)} \;\in\; \mathbb{R}^d, \tag{2}$$

where $\mathcal{L} = \{4, 10, 18, 24\}$ by default. A probe maps $\mathbf{h}_S$ to a truth score; we seek robustness of the probe's AUROC across the direct→reasoning shift $\mathcal{P}_{\text{direct}} \to \mathcal{P}_{\text{reason}}$.

A.10 Representation-shift analysis. We render the same KB facts as a direct statement and as a CoT statement that share an identical claim span. For each of the 28 layers we extract the claim-span hidden state and measure factuality separability (true vs false) in-style and cross-style, plus the direct↔CoT drift via claim-vector L2 distance and CKA (Kornblith et al., 2019). Factuality remains linearly separable in both styles at every layer (in-style AUROC roughly 0.89–0.99), so the signal is not destroyed. But the direct↔CoT CKA generally decreases with depth (with small reversals) to 0.46 at the top layer, and a direct-trained probe loses ≈ 0.09 AUROC on CoT ($0.992 \to 0.905$ at L26). CoT therefore induces a *family-dependent recoverable shift*; recalibration re-centers the classifier on the shifted representation. The three-panel figure for this analysis (Figure 1) appears in the main text (Section 5.3); the per-layer numbers behind panels (a)–(c) are tabulated in the corresponding block records.

TRUTH-LABEL-FREE, CLAIM-PAIR-SUPERVISED FACT-IDENTITY CONTROL (THE FACT-IDENTITY CONTROL). To show the drift is a *family-dependent recoverable shift* rather than noise, we pair the same claim across styles by its surface content (subj, rel, obj)—the map reconstructs the direct hidden vector and never sees the factuality label—then project both styles into the same 256-dim PCA basis and compare. The same-fact direct↔CoT claim-vector distance is **4.6**× smaller than the cross-fact distance (test: 5347 vs 24622); the two style subspaces share principal angles of essentially 0° and a matched-pair linear CKA of 0.825, so they occupy the same semantic space. A ridge-regularized linear map $T : \mathbf{h}^{\text{reas}} \to \mathbf{h}^{\text{dir}}$ (fit on training facts, width 32) further reduces the held-out same-fact distance by 37%. A label-shuffle control—scrambling which direct vector each CoT vector is mapped to—fails (3.5× larger residual), confirming the alignment reflects fact identity, not a spurious mean offset. The signal is shifted, and the shift is linearly recoverable in the mean (it is not a verified inverse).

PAIRED-INTERVENTION RECOVERY OF THE DIRECT-ONLY PROBE (THE PAIRED-RECOVERY PILOT). The key test is whether the truth-label-free, claim-pair-supervised map *restores probe performance*, not merely reduces distance. We train a direct-only probe in the shared 256-dim PCA subspace and evaluate on the held-out reasoning test under four mappings. In-style direct classification is the upper bound (0.971); the unmapped cross-style probe degrades (0.915). Applying the truth-label-free, claim-pair-supervised paired map T to the CoT states recovers AUROC to $\mathbf{0.982}$ ($+0.067$ over the unmapped probe, 95% CI $[+0.036, +0.102]$; ridge selected on a subject-disjoint validation split of the training pairs, test touched once). Two controls isolate the cause: a pure mean-shift (global offset) reaches only 0.920, and a shuffle-paired map collapses to 0.499 (chance; specificity $+0.483$, CI $[+0.416, +0.549]$). The recovery is therefore driven by fact-level structured shift that a claim-pair-supervised linear map can undo—the mechanism recalibration exploits.

SPECIFICITY OF THE RELOCATION (BLOCK 23). Two counterfactuals confirm the map recovers a *factuality-specific* structure, not merely an explicit relation-identity effect (run on Qwen-7B and Llama-3.1-8B, the two families where the paired map recovers). A key one is a *truth-composition transfer* test: a map fit on TRUE-only pairs recovers only 0.681 AUROC, and on FALSE-only pairs 0.734, while size-matched *mixed* pairs recover 0.981–0.985 (all-pairs 0.983)—the map needs both truth compositions, so it captures a truth-dependent structure, not a truth-agnostic style transform. A complementary probe-control is consistent: a relation-identity probe (10-way one-vs-rest) on the same states is already near-perfect under the shift (in-style A1 ≈0.999), so CoT realigns the factuality signal along a family/style-dependent direction—recovered by a claim-paired affine map on Qwen/Llama—rather than erasing it or transforming the explicit relation content (which the map recovers by only $+0.001$ vs. $+0.059$ for factuality on Qwen-7B). The shifted quantity is therefore the factuality signal, not the explicit relation identity/content.

GENERIC (PAIR-FREE) ALIGNMENT CANNOT RECOVER THE PROBE (BLOCK 30). A further control asks whether *unsupervised, distribution-level* alignment—the generic prompt-conditioned map family of Cheng & Kriegeskorte (2026)—suffices, which would reduce our paired map to an instance of generic representation alignment. On the same PCA subspaces and the same frozen direct-only probe we fit three pair-free linear alignments: mean-shift ($h \mapsto h - \mu_{\text{reas}} + \mu_{\text{dir}}$), CORAL (first/second-moment matching), and a PCA-component-matched orthogonal Procrustes (top-32 components paired by loading cosine, then Procrustes rotation). None recovers: on Qwen-7B they reach $0.925/0.842/0.652$ vs. 0.983 for the fact-paired ridge (shuffle control 0.465); on Llama-3.1-8B $0.933/0.887/0.366$ vs. 0.973; on Mistral-7B $0.949/0.891/0.483$ vs. 0.958; on Gemma-3-12B $0.744/0.744/0.511$ vs. 0.788. Mean-shift equals or under-performs the unmapped probe; CORAL and Procrustes are at or near chance. The recovery therefore requires *claim-level correspondence* (which claim in the CoT is the same fact as which direct answer), not style-level moments: the map is truth-label-free but pair-supervised, and the pairing is exactly what carries the signal.

PAIRED GAP VS. COT-REFERENCE OVER 5 SEEDS (BLOCK 29), WITH HOLM CORRECTION (BLOCK 35). The gap comparison extends both arms to 5 partition seeds under one protocol (same subject-disjoint splits, same probe family, same PCA). Per-family paired per-seed gaps (aligned − CoT-reference): Qwen-7B $+0.009 \pm 0.006$, 95% CI $[+0.002, +0.017]$, p=0.026, $4/5$ positive; Llama-3.1-8B $+0.025 \pm 0.010$, CI $[+0.012, +0.037]$, p=0.006, $5/5$; Mistral-7B $+0.003 \pm 0.007$, CI $[-0.006, +0.012]$, n.s.; Gemma-3-12B -0.115 ± 0.028, CI $[-0.149, -0.080]$, $0/5$. Under a Holm

correction across the 4 families: Llama p_{adj}=0.018 (significant), Qwen p_{adj}=0.053 (marginal), Mistral n.s., Gemma significantly *negative* (p_{adj}=0.003). On the two recovering families the aligned map reaches the frozen probe's own level (the map consumes the probe's direct truth labels, so it recovers—not surpasses—a truth-supervised probe); the gap's CI excludes zero.

PAIRED-MAP BASELINES, INVERTIBILITY DIAGNOSTICS, MULTI-PROBE TRANSFER, AND RELATION-DISJOINT RECOVERY (THE PAIRED-MAP CHARACTERIZATION). Four further experiments scope what the paired map is. *(a) Paired baselines* (same pairs, PCA features, frozen probe): paired orthogonal Procrustes reaches 0.956 (Qwen) / 0.944 (Llama), PLS-64 0.962/0.952, vs. the affine ridge 0.983/0.973; CCA-64 fails outright (0.458/0.439). So paired fitting is what matters (pair-free versions of the same methods fail, Block 30), and the full-rank affine capacity of the ridge contributes on top; CCA's canonical subspace alone is insufficient. *(b) Invertibility diagnostics:* the forward map W has effective rank $217/256$ (singular values $> 1\%$ of top), condition number 1.5×10^4; a reverse ridge map fit on the same pairs transfers a frozen CoT probe to mapped direct states at only 0.493 AUROC (chance), and the cycle $T_{\text{fit}}^{-1}\circ T$ reconstructs test states with 0.47 relative RMSSE. The map is therefore a probe-recovering alignment, *not* a verified bijection—which is why the paper claims linear *recoverability*, never invertibility. *(c) Multi-probe transfer:* the same frozen T recovers an independently trained linear probe ($0.827 \to 0.984$ Qwen, $0.942 \to 0.976$ Llama) and a centroid probe on Qwen ($0.816 \to 0.861$), but not Llama's centroid ($0.910 \to 0.875$): a coordinate change that transfers to some readouts, not all. *(d) Relation-disjoint:* re-fitting on 80% of relations (616 pairs) and evaluating only test items with *unseen* relations (54 items), the map still recovers on Qwen ($0.916 \to 0.955$) and holds on Llama ($0.952 \to 0.944$)—the transformation generalizes across relations, not just subjects.

LABEL-PRESERVING SHUFFLE: DECOMPOSING THE RECOVERY (BLOCK 71). The plain shuffle control breaks claim correspondence *and* class structure at once, so its collapse cannot separate the two. Block 71 shuffles the direct partner *within* each truth class (label marginals preserved, correspondence destroyed; 5 pairing resamples; ridge re-selected per family on the same val-MSE protocol as Block 9b, and the claim-paired arm reproduces Block 9b exactly: 0.982 Qwen, 0.974 Llama). Results: within-class shuffle reaches 0.972 ± 0.004 (Qwen) and 0.957 ± 0.005 (Llama), against plain-shuffle $0.472/0.479$ and unmapped $0.915/0.926$. The recovery has two components: class-conditional relocation ($+0.057/+0.031$ over unmapped – realigning CoT states to the label-conditioned manifold) and exact claim correspondence ($+0.015/+0.013$ on top, a paired subject-cluster bootstrap mean of per-resample differences; the difference of the displayed five-draw means is $+0.010/+0.017$ – the bootstrap pairs the full map against one shuffle draw rather than the mean, and both versions straddle zero: CIs $[-0.004, +0.050]/[-0.003, +0.035]$, Block 74). Two consequences. First, the paired estimator's advantage over pair-free fits (Block 30) is carried mostly by the class-conditional component – and the Block 74 logit distillation shows that component alone reaches within 0.005 of the full map; second, a fixed global transformation that permutes classes *consistently* would also be caught by this control, so the data cannot distinguish claim-specific geometry from a label-structured global transformation – the paper's claim is the decomposition and the necessity of paired fitting, not claim-specificity of the transformation.

SPONTANEOUS-COT MAP RECOVERY, POOLED ACROSS SEEDS (BLOCKS 37–38). On the fact-disjoint spontaneous-CoT test with *balanced* claim pairs (Block 37: the direct-side feature is generated for the exact (subject, relation, object) of every spontaneous claim, so pairs cover all ∼250 claims with their natural TRUE/FALSE mix instead of the ∼90%-TRUE overlap of a random direct-key set), the map recovers the frozen probe on 2 of 3 seeds (0.972/1.000/0.785, mean 0.919 ± 0.117; shuffle 0.44). The clean-span regeneration (Block 38, A.5g) sharpens the reading: its clean test set contains 2–21 FALSE items, so per-seed AUROC there measures ranking on a near-single-negative pool, and the pooled mean (0.737) is the quantity to quote for that condition. The mechanism's full-precision demonstration is on templated CoT, where the claim span is exactly localized, the false rate is controlled by construction, and recovery reaches 0.98 (Section 5.3).

A.11 Remaining ablations.

SHORTCUT CONTROLS (BLOCK 18). Four CPU-only baselines on the fact-disjoint reasoning test (Table 15)—text TF-IDF, object identity, triple identity, relation identity—are all at chance

(0.50–0.53), while hidden-state probes reach 0.88–0.99. The signal is internal, not a lexical or entity artifact.

Table 15: **Shortcut controls (Block 18, fact-disjoint reasoning test).**

Classifier	AUROC
text-only TF-IDF (1–2 gram)	0.530
object identity (one-hot)	0.497
(subj, rel, obj) identity (one-hot)	0.500
relation identity (one-hot)	0.515
last-layer-only hidden-state probe	0.880
full 4-layer logistic probe	0.911
trained MLP probe	≈0.99

PROBE CAPACITY (BLOCK 5A). Sweeping the probe head width from 32 to 512 hidden units, the gain is significant at every capacity (e.g., h=32: ΔAUROC $= +0.018$, CI $[+0.004, +0.036]$; h=256: $+0.023$, CI $[+0.007, +0.042]$). A 32-unit probe already captures the signal. Capacity is not the story.

TRAINING-SET SIZE (BLOCK 5C). The gain is unstable at very small supervision: at 150 labeled spans per style the gain is not yet significant (ΔAUROC $= -0.008$, 95% CI includes zero), whereas the grid already shows a positive significant gain at 600 per style ($+0.022$, CI excludes zero) and $+0.024$ at 1200 per style.

WHERE THE SIGNAL LIVES (BLOCK 5E). Re-extracting all 28 layers, single-layer AUROC peaks in the middle (L17 = 0.992; L0 = 0.905). The default quad $\{4, 10, 18, 24\}$ (0.990) is already near the all-28-layer probe (0.991); the top-3 middle layers alone match it (0.991). The cue is concentrated in mid layers and redundant.

Table 16: **Ablation summary (Block 5): training composition (direct:reasoning).** The gain tracks reasoning exposure, not data volume. All AUROCs on the reasoning test set.

ratio	ΔAUROC	95% CI	holds
1:0	-0.012	$[-0.022, -0.004]$	✗
1:0.25	+0.019	$[+0.004, +0.037]$	✓
1:0.5	+0.021	$[+0.005, +0.039]$	✓
1:1	+0.016	$[+0.004, +0.031]$	✓
1:2	+0.017	$[-0.001, +0.037]$	∼

KEYSTONE PER-SEED VARIANCE (BLOCK 2.5). The keystone baseline AUROC varies across probe-initialization seeds because the direct-only probe is trained on a small, budget-starved spontaneous-proxy slice and the test set is the model's own unbalanced free-form proxy generation. Table 19 reports the per-seed baseline AUROC for each family so the variance is visible, not just summarized as mean±std. The recalibrated probe is near ceiling on Qwen, Mistral, and Llama (0.9886 or higher in every seed there; one Mistral seed is exactly 0.9886), whereas Gemma is lower (≥ 0.89), so its low std reflects saturation rather than a stability claim; the baseline variance is large exactly because the baseline is not at ceiling. The headline conclusion—that target-style recalibration gives high AUROC on Qwen/Mistral/Llama and substantial recovery on Gemma—is robust to the seed (3/3 seeds for Qwen, Mistral, Llama; the Gemma mixed probe reaches 0.89/0.96/0.95 across seeds, all ≥ 0.89).

Note on the Gemma seed-1 baseline (0.500). The chance-level AUROC is the budget-starvation regime the keystone is designed to expose: the recalibrated probe trained on the *same* Gemma seed-1 spontaneous set reaches 0.892 (Gemma mixed-probe seeds $0.892/0.965/0.943$), so the labels and the hidden states carry the signal, and the direct-only baseline locates it once target-style exposure enters the budget.

Table 17: **Equal-budget composition matrix (Block 5F).** Every config trains on the same total B=2400 labeled spans; only the direct:reasoning split varies. Doubling direct supervision alone (2400:0) does *not* recover the gain (ΔAUROC vs direct-only $= -0.024$, CI excludes 0), while adding even 480 reasoning spans (1920:480) restores it ($+0.016$, CI excludes 0). Only configs containing reasoning reach high AUROC on the reasoning test, and AUROC on the direct test stays ≥ 0.97 throughout. Reasoning exposure, not data volume, is the active ingredient.

Config (direct:reasoning)	B	$\text{AUROC}_{\text{reas}}$	$\text{AUROC}_{\text{dir}}$	ΔAUROC vs direct-only (95% CI)
2400:0 (2× direct)	2400	0.943	0.989	-0.024 $[-0.035, -0.014]$
1920:480 (1:0.25)	2400	0.982	0.985	$+0.016$ $[+0.001, +0.033]$
1600:800 (1:0.5)	2400	0.987	0.987	$+0.021$ $[+0.006, +0.039]$
1200:1200 (1:1)	2400	0.991	0.987	$+0.024$ $[+0.009, +0.043]$
800:1600 (1:2)	2400	0.984	0.983	$+0.017$ $[+0.001, +0.036]$
480:1920 (0.25:1)	2400	0.984	0.985	$+0.018$ $[+0.004, +0.034]$
0:2400 (reasoning-only)	2400	0.979	0.972	$+0.013$ $[-0.002, +0.031]$

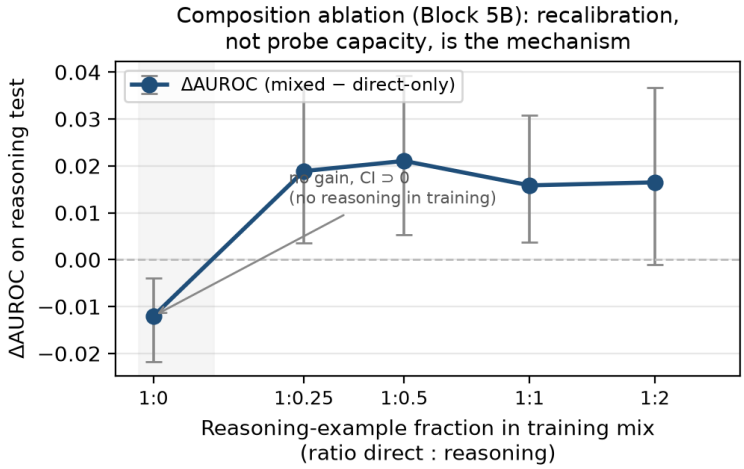


Figure 3: **Composition ablation (Block 5B).** Without reasoning examples in training (1:0) the probe collapses to the direct-only baseline (ΔAUROC CI excludes 0 on the negative side—no gain); adding reasoning recovers the gain. The training *distribution*, not the probe, is the mechanism. Generated from `block5_ablation.json` (Block 5B, fact-disjoint).

REPRODUCIBILITY. The clean regeneration prompts (neutral and forced `Answer:` formats), the last-vocabulary span rule, and the regex/LLM self-localizers used in Appendices A.5g/A.12 are fully specified in the corresponding block descriptions in this appendix.

A.12 Localization and multi-claim evaluation (Blocks 42–44).

LOCALIZATION METRICS (BLOCK 42). The deployment bottleneck is claim-span localization, not probe quality. On PopQA (Qwen, 3 seeds) the fully paired gold-span aggregate is AUROC 0.780 ± 0.017, false-AUPRC 0.872 ± 0.036, ECE 0.274 ± 0.041, and FPR@95TPR 0.717 ± 0.025. The only seed with both cached gold and regex results (seed 20370101) gives $0.799/0.922/0.330/0.721 for gold spans and 0.582/0.823/0.312/0.814$ for the deployment-style auto-localizer, an AUROC change of -0.217. Gold recall is only 0.22–0.34 and claim-match F1 0.04–0.08 (regex Block 42); Mistral's gold-span F1 is higher (0.46–0.53), but no Mistral auto-span AUROC is cached. The conclusion is unchanged: localization—not the probe— dominates measured end-to-end performance.

LLM SELF-LOCALIZATION (BLOCK 44). The regex localizer's failure is a *front-end* problem (claim-match F1 0.04–0.08), so we ask whether a zero-label LLM can localize its *own* asserted span. We show each model its own cached CoT and prompt it to copy verbatim the single asserted-answer

Table 18: **Supervision-aligned baselines on the reasoning test (Block 15).** All methods score the same v3-disjoint reasoning test; AUROC (higher is better). Platt/isotonic and CORAL/mean-shift are unsupervised or calibrator-only; the paired ridge map (Block 9b) is truth-label-free at alignment time. The mixed probe and the paired ridge map both recover near the direct-only probe's own in-style AUROC.

Method	Alignment / calibration	$\text{AUROC}_{\text{reas}}$
Direct-only probe (frozen)	none	0.933
+ Platt scaling (held-out fit)	1-D monotone calibrator	0.933
+ isotonic (held-out fit)	1-D monotone calibrator	0.929
+ CORAL (reasoning→direct)	unsupervised global cov.	0.500
+ mean-shift (global offset)	unsupervised global mean	0.929
+ paired ridge map (Block 9b)	truth-label-free fact-aligned	**0.982**
Mixed probe (direct+reasoning)	target-distribution train	**0.983**

Table 19: **Keystone per-seed baseline AUROC (Block 2.5 proxy).** The direct-only baseline on the free-form proxy is seed-variable (it is trained on a small slice of an unbalanced proxy set); the recalibrated probe is near ceiling on Qwen/Mistral/Llama but lower on Gemma. This table makes the variance behind the mean±std visible.

Family	seed 1	seed 2	seed 3	mean±std
Qwen-7B	0.704	0.793	0.914	0.804 ± 0.105
Mistral-7B	0.823	0.776	0.890	0.829 ± 0.057
Llama-3.1-8B	0.877	0.876	0.921	0.891 ± 0.026
Gemma-3-12B	0.500	0.559	0.636	0.565 ± 0.068

span; that span is wrapped and scored with the *identical* probe+PCA as Block 26, so the only variable is localization. On Qwen2.5-7B (3 seeds, $n{\approx}195$) the LLM localizer raises probe AUROC from the regex's 0.582 (seed 01) to 0.685 (seed 01 LLM, $+0.103$) and a 3-seed mean of $0.705{\pm}0.044$ (oracle $0.780{\pm}0.017$); the mean residual (LLM − oracle) -0.075 is within seed noise (paired $t{=}2.18$, $\text{df}{=}2$, n.s.). Because the gold oracle is recall-limited—at one Qwen seed it falls below the LLM localizer (-0.009)—-0.075 is an upper bound on the LLM's span-copy shortfall, not a deficit below a hard ceiling. The remaining gap is the LLM's own imperfect span copy: on Qwen it *finds* the asserted span ≈ 0.90 of the time but copies it exactly only ≈ 0.18 (claim-match F1)—the exact-copy shortfall is the bottleneck, not probe quality—confirming (iii): the deployment bottleneck is localization, and a free, label-free LLM front-end recovers roughly half of the paired seed's oracle-to-regex gap on Qwen (it adds one LLM forward pass per claim; on Mistral it matches the regex localizer, see Table 20). On Mistral-7B the oracle spans themselves span 0.556–0.859 across seeds (mean AUROC $0.686{\pm}0.127$), which sets the reference band the LLM localizer ($0.609{\pm}0.037$) lands inside.

MULTI-CLAIM PILOT (BLOCK 43). We test the W2 concern—does per-claim scoring survive when *several* claims share one CoT trace?—with a controlled paired design on Qwen-7B (3 seeds). One free-form CoT asserts several facts about a single subject (capital / language / currency / largest-city / continent of a country); we localize each claim's span and, for every gold claim the model states, insert a *false twin* (the value swapped to a different object in the same relation) to guarantee class balance and isolate localization+scoring under multi-claim context (spontaneous natural errors also kept). The direct-only probe flags the false twins at AUROC 0.991 / 0.987 / 0.988 (mean 0.989), every attribute ≥ 0.97. This shows per-claim localization+scoring holds when multiple claims co-occur in one trace. The pilot is scoped as *single-subject, multi-attribute*: the synthetic KB has disjoint subject/object types, so it cannot express chained hops. The multi-hop measurement is the corrected PrOntoQA step audit (Blocks 67/69), which quantifies the natural step error rate (0.5%) rather than a step-level probe; knowledge-grounded multi-hop corpora with unverifiable free-text claims (HotpotQA / 2Wiki style) remain outside our setup.

Table 20: **Localization ablations on PopQA (Blocks 42/44).** Probe AUROC under three span conditions: oracle gold-span (Block 26 upper bound), deployment regex (Block 42, Qwen seed 01), and zero-label LLM self-localization (Block 44, 3 seeds). The only variable is localization; the probe+PCA are identical. Regex claim-match F1 0.04–0.08 is from Block 42; the LLM-localizer F1 ≈ 0.18 is from Block 44. On Mistral the LLM localizer reaches F1 ≈ 0.23 vs the regex ≈ 0.50, so it does not beat regex—the roughly-half-gap result is Qwen-specific to the paired seed.

Condition	Qwen AUROC	Mistral AUROC	claim-match F1
Oracle (gold-span)	0.780 ± 0.017	0.686 ± 0.127	–
Regex (deployed)	0.582	–	0.04–0.08
LLM self-localize	0.705 ± 0.044	0.609 ± 0.037	≈ 0.18

Table 21: **Multi-claim pilot (Block 43, Qwen-7B, controlled paired design, 3 seeds).** AUROC of the direct-only probe on detecting false claims embedded among true ones within one multi-claim CoT. Base rate is approximately 0.5 by the false-twin insertion.

Seed	n_{claims}	true	false-twin	AUROC
20370101	422	211	211	0.991
20370102	411	202	202	0.987
20370103	410	205	205	0.988

A.13 Data and protocol audit. To make the experimental lineage auditable we list, per main claim, the model(s), data source, probe-training composition (and the reasoning style it contains), train/test disjointness, how claim spans were obtained, and whether gold labels are used at inference.

Table 22: **Data and protocol audit.** "mixed" probe = direct + reasoning-style examples. KB-QA uses a 125-triple synthetic knowledge base expanded by templates; v3-disjoint = subject-disjoint split.

Claim / table	Model(s)	Data source	Probe train (reasoning style)	Split / span / gold
C1 (Table 9)	Qwen 1.5B/7B, Mistral-7B, Llama-8B, Gemma-3-12B	125-triple KB-QA, templated direct + templated CoT	baseline 1200 direct; mixed 1200 direct + 1200 templated CoT	subject-disjoint (v3); `<OBJ>` span; no gold@inference
Keystone (Table 4)	same 4 families	model's free-form CoT, vocabulary-order asserted-object proxy labels (no answer injection)	baseline = templated direct; mixed = templated direct + balanced *spontaneous-proxy* CoT (train split)	prompt-disjoint spontaneous train/test; proxy-false rate ≈19%; heuristic asserted span; no gold@inference
Keystone fact-disjoint (Block 2.5b)	Qwen-7B	same KB-QA, subject-disjoint free-form CoT generation	baseline = direct (templated); mixed = direct + templated CoT + spontaneous-proxy CoT	**subject-disjoint** train/test (3 partition seeds); no subject overlap; proxy-false rate ≈15–20%; heuristic asserted span; no gold@inference
Composition / equal-budget (Figure 3, Table 17, Block 5F)	Qwen-7B	same KB-QA	composition ratios 1:0–1:2 plus fixed-B=2400 matrix 2400:0, 1920:480, 1600:800, 1200:1200, 800:1600, 480:1920, 0:2400 (2:0 and 0:1 realized); length controlled via shared `<OBJ>` span	subject-disjoint (v3); same as C1
Real corpora (Table 12, app)	Qwen-7B, Mistral-7B	PopQA, TruthfulQA, TriviaQA	synthetic probe, then corpus-specific in-distribution labels (PopQA 190–200 retained; TruthfulQA 44; TriviaQA 234)	corpus test; gold / last-mention span; gold used only for oracle span localization, not probe scoring at deployment

Table 22 (continued)

Claim / table	Model(s)	Data source	Probe train (reasoning style)	Split / span / gold
Baselines (app)	all	as above	pTrue (verbalized), Semantic Entropy	same test sets; no probe training
Representation shift / mechanism (A.10, Block 9/9b)	Qwen-7B	same KB-QA, same-fact content pairs	truth-label-free; direct-only probe; paired ridge map	subject-disjoint (fit/val selects ridge); PCA 256-d; test touched once; no gold@inference
Supervision-aligned baselines (Block 15, Table 18)	Qwen-7B	same KB-QA, v3-disjoint reasoning test	frozen direct probe + Platt/isotonic (held-out reasoning-train fit); CORAL; mean-shift; paired ridge map (truth-label-free)	subject-disjoint (v3); calibrators never touch test; no gold@inference
Unified factorial (Table 1, Block 64)	Qwen-7B	125-triple KB-QA, v3-disjoint caches	logistic probe (L2, C=1, max_iter 4000, z-scored inputs; 5 probe seeds/cell); arms direct-only / target-only / mixed within same total budget $B \in \{60, \ldots, 960\}$	PCA-256 fit on pooled *train* direct+reasoning rows per cell; subject-disjoint test touched once. PCA bases are fit on label-free feature rows (8,000 total, both styles) – this uses no truth labels and is reported separately from the label budget B (Block 72b reruns key cells with the MLP head)
Label-preserving shuffle (A.10, Block 71)	Qwen-7B, Llama-3.1-8B	same KB-QA caches as Block 9b	block9b protocol (MLP probe, PCA-256, per-family val-MSE ridge: $10^4/10^2$); within-class partner permutation, 5 resamples	A3 reproduces Block 9b exactly (0.982/0.974)
Three-way adjudication (A.7, Block 72)	Qwen-7B, Mistral-7B	all cached PopQA natural-label generations (1,185 test items, 3 seeds)	7 terminal-assertion regex families + alias-table entity match; probe retrained on adjudicated-only labels	old labels 0.993 mean $\rightarrow$ adjudicated 0.932; 27.5% of auto-FALSE are asserted errors
Readout-head sensitivity (A.5e, Block 72b)	Qwen-7B	Block-64 factorial cells (raw/PCA $\times$ direct/target, $B \in \{60, 240, 960\}$)	MLP-256 head (AdamW) vs. logistic on identical caches	target-only beats direct-only $6/6$ under both heads; levels shift -0.11 to $+0.02$
Per-unit counts + cluster CIs (A.7, Block 73)	Qwen-7B, Mistral-7B	Block-72 adjudicated subsets	question-level cluster bootstrap (2,000 resamples), train fixed; per-unit class counts	retrain 0.817–0.995 (CIs per unit); frozen alias probe 0.958–1.000
Logit distillation + increment CIs (A.10, Block 74)	Qwen-7B, Llama-3.1-8B	same pairs/splits/probe as Block 71	same-pairs ridge to frozen-probe logit (scalar, label-free); subject-cluster bootstrap of A3−A5, A3−D	distill within 0.005 of full map; exact-correspondence increments $+0.013$–0.015, CIs straddle zero
Same-assertion pairing (A.8, Blocks 75/75b)	Qwen-7B	PopQA pair caches of Block 21b	paired ridge map, direct side renders the model's asserted object (no answer key, no alias table); 5 partition seeds, PCA pool fixed to Table-2 protocol; dual-pool decomposition run separately	same-assertion 0.953 ± 0.010 vs. gold-conditioned 0.840 ± 0.038; shuffle 0.527 ± 0.049; logit distill 0.950 ± 0.004; $5/5$ sig
Independent LLM verification (A.7, Block 76)	Qwen-7B (judge), Qwen/Mistral traces	all 2,400 kept natural-label generations	zero-shot A/B/C verdicts (question + accepted answers + trace only); probe retrained on judge semantic labels	cue survives: 0.981/0.985 judge-label retrain; 84% of regex-FALSE are truncated per judge

Table 22 (continued)

Claim / table	Model(s)	Data source	Probe train (reasoning style)	Split / span / gold
Same-instance detector baselines (A.7, Block 77)	Qwen-7B, Mistral-7B	Block-72 adjudicated test items	probe retrained / frozen alias / p(True) / calibrated span-entropy on identical test items	retrained 0.820–0.995; frozen alias 0.960–1.000; p(True) 0.430–0.527; span-entropy 0.437–0.885

Protocol notes (factual scope). (i) For C1 probe-init rows, the three reported seeds vary only probe initialization and training; the underlying generations and KB split are fixed, so they measure training stability, not fact- or prompt-sampling variance. Other blocks (for example the feature factorial and paired-map diagnostics) have their own partition/train–validation semantics recorded in the corresponding tables. (ii) The keystone mixed probe is trained on the target (spontaneous) distribution and evaluated on a disjoint spontaneous split, so the $0.804 \rightarrow 0.999$ result is target-domain recalibration (no test contamination), not zero-shot templated→spontaneous transfer; moreover the 0.804 baseline is *budget-starved* (a roughly $2\times$ supervision imbalance; per-arm counts were not logged), and the proxy labels themselves were retracted by the Block 55 audit – the row is retained as measurement history only, with no bound interpretation and no role in current factual claims (see Blocks 2.5c/2.5d, Table 5). (iii) Core controlled probes use grounded claim spans; automatic localization is evaluated separately (Blocks 42–44).

A.14 Span-confound factorial probe (Block 45). The paired map's recovery is dominated by class-conditional label transfer with a small, statistically unresolved exact-correspondence increment (Blocks 71/74, Appendix A.10), but a reader may ask whether it recovers *factuality* or merely exploits the *exact-object span*. We separate the two with a factorial probe on the clean-spontaneous protocol (Mistral-7B): for each held-out claim we build *three* paired-direct targets that hold claim identity, relation, and span location fixed and vary *only* the wrapped object—(correct) the true answer object, (wrong) a different grammatical object of the *same* relation, (shuffle) a random claim→object pairing. We fit three affine maps on the same spontaneous train pairs, each to one target, apply each to the held-out spontaneous test states, and measure (i) reconstruction MSE to the respective target (base-rate-independent, full 243-claim test set) and (ii) probe AUROC on the test claims.

Table 23: **Span-confound factorial (Block 45, Mistral-7B, 3 seeds).** Reconstruction MSE of the mapped spontaneous states to each paired-direct target; lower means the map recovers that target. The correct-object target is recovered 3–6$\times$ better than a wrong-object or shuffle target, so the map exploits the *exact* object span, not a claim-identity-agnostic cue. Probe AUROC is high-variance on this near-degenerate pool (4–6 false claims/seed) and serves only as a consistency check.

Target	seed 20370101	seed 20370102	seed 20370103	mean $\times$ vs correct
correct-object	33.0	42.0	32.8	1.0$\times$
wrong-object	112.5	111.4	117.4	2.65–3.58$\times$
shuffle	198.0	197.7	197.8	4.7–6.0$\times$

The reconstruction error to the *correct* object is 3–6$\times$ smaller than to a wrong or shuffled object (Table 23), and the gap is stable across all three seeds (mean wrong/correct $= 3.2\times$, shuffle/correct $= 5.6\times$). Therefore the map's recovery is *mediated by the exact-object span*: swapping only the wrapped object (same claim, same relation, same span location) inflates reconstruction error by 3–6$\times$, and a shuffle target is worst of all. Probe AUROC shows no clean ordering (the clean-spontaneous pool contains only 4–6 false claims per seed, so the arm is high-variance and uninformative); we rely on reconstruction MSE. *Conclusion (audit finding):* the recovery map is partly a span-confound exploit—it recovers factuality on the in-domain exact-span protocol but does *not* isolate a claim-identity-agnostic mechanism. This sharpens the controlled-audit scope and motivates the direct-only / target-only recommendation (Section 5.2), which does not depend on the map.

A.15 Object-disjoint map recovery (Block 46). The span-confound factorial (A.14) shows the map is keyed to the exact-object span, but leaves open a stronger alternative: does the map merely *memorize*

object-token correspondences? We isolate this on the templated protocol with an *object-disjoint* axis (Qwen-7B and Llama-3.1-8B, the two families whose map recovers the probe; seeds 20370101–3, identical probe/PCA/ridge protocol to A.10): we fit the claim-paired ridge map on training pairs whose object vocabulary covers only 54 of 68 objects, then apply it to held-out test claims whose object was *never* seen in map training (248 claims; relations still overlap train, so the axis is object identity, not relation coverage). The map recovers the probe on test claims whose objects it *never saw* in training

Table 24: **Object-disjoint recovery (Block 46).** The claim-paired map is trained on object-vocabulary-disjoint training pairs and applied to held-out test claims. "object-disjoint" = test object never in map training (the key arm); "same-object/diff-fact" = test object was in training but its (subject, relation) fact is new; "in-domain" = all pairs (sanity, A.10); "shuffle" = shuffle-paired control. AUROC of the frozen direct-only probe on the mapped CoT states.

Family	Arm	Mapped	Unmapped	Δ	n
Qwen-7B	in-domain (sanity)	0.983	0.924	$+0.059$	300
Qwen-7B	object-disjoint	0.980	0.917	$+0.063$	248
Qwen-7B	same-obj/diff-fact	0.994	0.927	$+0.067$	52
Qwen-7B	shuffle control	0.546	—	—	300
Llama-8B	in-domain (sanity)	0.973	0.929	$+0.044$	300
Llama-8B	object-disjoint	0.971	0.923	$+0.048$	248
Llama-8B	same-obj/diff-fact	0.984	0.959	$+0.026$	52
Llama-8B	shuffle control	0.607	—	—	300

(seed 20370101: Qwen $0.917 \rightarrow 0.980$, $+0.063$; Llama $0.923 \rightarrow 0.971$, $+0.048$), essentially matching in-domain recovery. Two further seeds (20370102, 20370103; same CPU-only protocol) confirm the object-disjoint gap on every family/seed ($6/6$ positive): Qwen unmapped→mapped $0.915 \rightarrow 0.969$ and $0.882 \rightarrow 0.960$, Llama $0.904 \rightarrow 0.961$ and $0.894 \rightarrow 0.950$—so the paired map is *not* an object-token memory. Same-object/different-fact transfer is at least as good (0.994/0.984), so object identity adds a small cue but is not necessary. Together with A.14 (swapping the wrapped object on the *same* claim inflates reconstruction error 3–6×), the map's true nature is a *content-paired, exact-span representation alignment*: for each claim it fits the corresponding direct representation tightly (span-sensitive), yet it generalizes along claim content rather than memorizing object tokens (object-disjoint holds). This bounds the mechanism claim to *exact-span paired representation alignment*—it does not assert a universal factuality geometry, and it remains specific to the two families where the frozen probe is linearly recoverable.

A.15a Pair cost versus label cost (Block 34). The claim-paired map needs no truth labels, but pairs are not cheaper than labels. On the templated protocol (Qwen-7B, 3 partition seeds) we sweep both acquisition modes against each seed's direct-only baseline at $B{=}120$ (0.872–0.943): 24 truth-labeled target-style examples already match the baseline on every seed, whereas matched-pair acquisition is unstable—one seed needs $\sim$150 pairs to exceed it, a second reaches it at 20 but dips below until $\sim$93, and the third never stably exceeds it within its pair pool (205 pairs). Per unit of annotation effort, truth labels therefore dominate paired claims for recovery, which is why we present the map as a mechanism probe rather than a deployment route.

A.16 Three map classes fail to invert Gemma's relocation (Block 24/25). Gemma-3-12B's linear ridge map fails ($0.786 \rightarrow 0.777$, n.s.). We tested two additional map classes on the same fact-aligned pairs and subject-disjoint split as Block 19b: *(a)* a 2-layer MLP map (256 hidden, ReLU, 200 epochs, AdamW), reaching $0.794{\pm}0.024$ ($0/3$ seeds significant, shuffle control 0.45–0.57, Block 24); *(b)* a k-NN retrieval map (cosine similarity on the fact-aligned train pairs, swept $k \in \{1, 5, 10\}$ over 5 seeds), reaching best $k{=}10$: $0.790{\pm}0.019$ ($0/5$ seeds significant, Block 25). All three map classes fall short of the in-style CoT reference ($0.878{\pm}0.013$, the 3-seed value used throughout Table 2; a single-seed diagnosis run reports 0.872): on Gemma the relocation is reachable only by a target-style probe among the function classes tested on fact-aligned pairs. Gemma's structured residual (aligned−shuffle, per seed) *is* positive and consistent ($+0.300{\pm}0.029$, $3/3$ seeds), so the map isolates a fact-level signal on Gemma as well – one whose size leaves the gap to a directly-supervised in-style probe to the label route (Section 5.2).

A.17 Protocol factorial: prompt, length, and span effects (Block 47). The clean-regeneration comparison in the spontaneous-proxy protocol (A.5/A.5g) changed the prompt format, the generation budget, and the span rule at once. To attribute the 15–20% → 3–7% false-claim-rate drop we run a 2×2 factorial on the *same* facts and seeds (Qwen2.5-7B, 3 seeds): prompt $\{$neutral, `Answer:`$\}\times$ $\max_{\text{new}}\{160, 512\}$. Cell A (neutral/160, original keystone data) and cell D (`Answer:`/512, clean-pool data) are the two previously reported protocols; cells B (neutral/512) and C (`Answer:`/160) are newly generated (Block 47). Per-cell metrics are pooled over train+test, mean±std over seeds; truncation is measured at generation time for cells B/C and by end-of-generation heuristics for the cached cells A/D; the false-claim label is the relation-scoped last-object rule (neutral cells) or the `Answer:`-line object (cells C/D).

Table 25: **Protocol factorial (Block 47, Qwen2.5-7B, 3 seeds).** Cell sizes and retention are as recorded in the raw artifact. The descriptive prompt main effect is -11.9pp versus a length main effect of -0.7pp; no inferential test for this contrast is stored in the artifact.

Cell	prompt	$\max_{\text{new}}$	trunc. rate	label. rate	false-claim rate
A	neutral	160	0.625±0.027	0.846±0.040	0.165±0.010
B	neutral	512	0.044±0.027	0.904±0.050	0.145±0.014
C	`Answer:`	160	0.960±0.013	0.779±0.009	0.033±0.011
D	`Answer:`	512	0.000±0.000	0.886±0.009	0.040±0.014

Two facts stand out. First, descriptively, the *length* factor moves the false-claim rate little (cell B is 95% untruncated yet still 14.5% inaccurate; cell C is 96% truncated yet only 3.3% inaccurate), while the *prompt* factor moves it by -11.9pp: the forced `Answer:` format coincides with a large descriptive change in the measured rate; behavior versus measurement is not fully isolated. Second, the span rule is a minor contributor: on the `Answer:` cells the crude last-object rule and the `Answer:`-line rule disagree on only 10–12% of items (span-agreement 0.877–0.896). For the neutral cells the false rate uses the same relation-scoped rule as the original 15–20% figure, so cells A/B are rule-consistent, though the rule's fidelity on free-form CoT is not separately measured. Cell C is 22% unlabeled (truncated before the `Answer:` line), so its 3.3% rate is conditional on labelability. The measured false-claim base rate is therefore *protocol-dependent*: ≈ 15% under the neutral prompt (even untruncated) vs. 3–7% under the forced format; the screening cost of ∼1,400–3,300 generations per 100 false claims on the clean spontaneous pool (A.5d) is the clean-pool (cell-D) rate.

Probe transfer is protocol-robust (Block 48). The ≈0.9 direct-probe transfer on spontaneous CoT could be an artifact of the forced `Answer:` protocol. We extract hidden states at the asserted-answer span for cell B (neutral/512, 3 seeds) and score the same direct-trained probe (balanced 120 direct labels, PCA as in Block 38) on all three spontaneous protocols: AUROC 0.914±0.037 (cell A, neutral/160), 0.844±0.089 (cell B, neutral/512), 0.928±0.066 (cell D, `Answer:`/512, reproducing the clean-pool number exactly). The transfer holds under every protocol, and the neutral cells' higher false-claim base rate (≈ 13–16% pooled, 12–20% per seed; 29–52 negatives per seed) makes their AUROC *less* degenerate than the clean pool's 2–21 negatives; cell B's gap to the clean pool (-0.08; cell A ≈-0.01) is a protocol-level difference, not a failure to transfer.

Second family: Mistral-7B (Block 49). To check that the protocol-robust transfer is not Qwen-specific, we repeat the neutral/512 protocol on Mistral-7B (3 seeds, same KB prompts, same last-object extraction rule). The neutral false-claim rate is 8.8%±0.9% (lower than Qwen's 13–16% pooled) and the same direct-trained probe (balanced 120 labels from Mistral's own direct pool) transfers at AUROC 0.809±0.042 (0.839/0.826/ 0.761 per seed, 20–30 negatives per seed). The ≈0.8–0.9 spontaneous transfer therefore holds on a second family, at a lower base rate than Qwen's neutral cells.

A.18 Cross-layer dynamic probe: architecture is not the active ingredient (Block 52). A stronger probe architecture could in principle absorb the style shift itself. ICR-style methods (Zhang et al., 2025) argue that *cross-layer dynamics*—not static representations—carry the hallucination signal. We therefore test, under the exact direct-train→CoT-test protocol of the main text (Qwen2.5-7B-Instruct, fact-disjoint split, data seed 20370101, 3 training seeds), five probe feature sets built from the same cached all-28-layer claim-span hidden states (no new generation): *static-4L* (the paper's static concatenation, layers $\{4, 10, 18, 24\}$), *icr-diff-4L* and *icr-diff-27* (adjacent-layer dif-

ferences $h_{l+1}-h_l$ at the same slots / all 27 layers, in the spirit of the ICR score, which additionally uses attention projections we do not cache), *icr-norm-27* (per-layer $\|h_{l+1}-h_l\|_2$ scalars), and the static+diff concatenation. Table 26 reports direct-only training, tested on held-out direct and CoT splits.

Table 26: **Cross-layer dynamic probe does not remove the CoT shift (Block 52, Qwen-7B, 3 seeds).** Direct-only training; AUROC mean over seeds; "drop" = direct-test − CoT-test.

Feature set	AUROC (direct test)	AUROC (CoT test)	drop
static-4L (paper baseline)	0.995	0.960	+0.035
icr-diff-4L	0.995	0.959	+0.035
icr-diff-27 (all layers)	0.989	0.966	+0.023
icr-norm-27 (scalar norms)	0.809	0.697	+0.112
static-4L + icr-diff-4L	0.995	0.961	+0.033

The full cross-layer dynamic feature transfers *as well as* the static probe (0.966 vs. 0.960) and its residual CoT drop is only modestly smaller (0.023 vs. 0.035); the scalar-norm caricature collapses (0.697), consistent with the signal living in the *direction* of layer updates rather than their magnitude. Adding dynamic features to static ones changes little (0.961). Under mixed training the vector-valued feature sets recover to roughly 0.985–0.993, whereas the scalar-norm baseline remains at 0.766–0.787. We conclude that *probe architecture—static, cross-layer dynamic, or their union—does not remove the style shift*; target-style exposure does. This bounds the "architecture is not the active ingredient" claim to the feature families tested here (concatenation and adjacent-layer dynamics); trajectory-shaping methods that intervene during decoding (e.g. ARS) are a different axis and remain complementary.