ANSWERING BY FORMAT: HOW EVIDENCE FORMATTING SHAPES LLM ABSTENTION IN RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

We study evidence-grounded answering and abstention in retrieval-augmented generation (RAG), where the decision to answer should depend on whether the evidence is sufficient. However, we identify a *spurious evidence-rendering gate*: changing how the same evidence is presented, without altering its content or order, can switch a large language model (LLM) between answering and abstaining. In controlled experiments with nine LLMs, correlating evidence format with whether training examples call for an answer or abstention induces this gate in eight of them. On insufficient evidence, this training increases the answer-rate gap between the two formats by an average of 0.70. Reversing this association reverses the learned format bias in most models. To assess the performance consequences of this association, we vary its strength and find that accuracy on the answer-or-abstain task declines as the association strengthens. To examine whether such correlations also occur in practice, we audit public training corpora and find associations between formatting cues and answer-versus-abstention labels. Controlled retraining further shows that artificially introducing such correlations can induce the gate. Beyond these controlled experiments, we also observe format-dependent answering in open (ChatQA-1.5-8B; TrustAlign-Qwen2.5-3B/7B) and closed (claude-opus-5) models. In the controlled experiments, removing the association between evidence format and answer-versus-abstention labels improves task accuracy by up to 0.37 relative to format-correlated training. Reliable abstention in RAG therefore requires controlling evidence presentation during both training and evaluation. We release our code at `https://anonymous.4open.science/r/paper-artifact-staging-B0B7/`.

1 INTRODUCTION

Retrieval-augmented generation (RAG) turns answering into an epistemic decision: an LLM should answer when the retrieved evidence is sufficient and abstain when it is not (Joren et al., 2025; Song et al., 2025). Under the complete-groundedness standard of grounded RAG evaluation (Song et al., 2025), the model must answer from the provided evidence alone: when the evidence cannot support an answer, the correct action is abstention, and any substantive answer is an unsupported assertion — even one that parametric knowledge happens to make factually right. This paper shows that the decision has a second, unguarded input: answer-versus-abstain behavior can be gated by how the evidence is *rendered* — the surface formatting wrapped around the retrieved items — rather than by what the evidence contains. We call this a *spurious evidence-rendering gate*. Concretely, we compare two content-fixed renderings of the same evidence: *chain* wraps the ordered items in a step-by-step scaffold, *raw* concatenates them plainly; content and order are byte-identical. The gate is easy to introduce and hard to see: abstention tuning draws its supervision from data in which the rendering happens to correlate with the answer/abstain label, a model can learn to read the rendering as the signal, and because the correlation lives in the surface form, an evaluation that matches evidence content but leaves its rendering uncontrolled does not expose it.

Why this matters: the gate is already present in released RAG models. The stakes are highest where RAG models work over specialized documents — ContractNLI (Koreeda & Manning, 2021), whether a contract entails a legal claim; ShARC (Saeidi et al., 2018), questions under the regulations governing government services and benefits — where a model that answers on insufficient evidence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

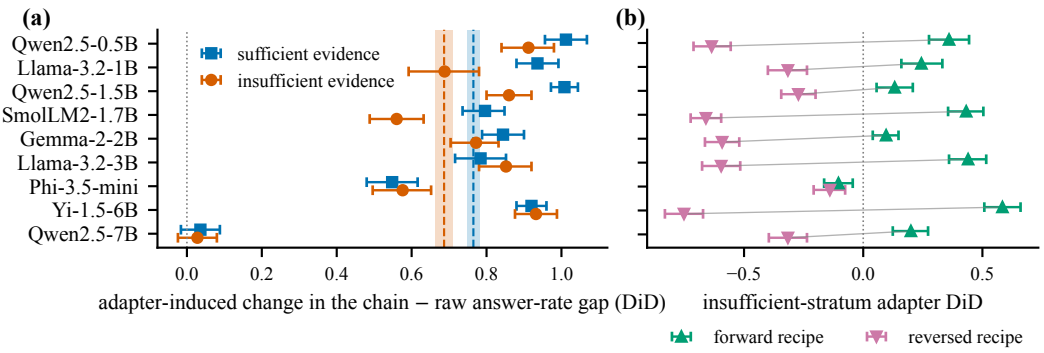

Figure 1: A content-fixed rendering gate under format-correlated abstention tuning (250 2Wiki development questions; semantic reading of every response, Section L; models ordered by parameter count). **(a)** Recipe A, deterministic re-run at seed 7: per-model adapter DiDs in the chain−raw answer-rate gap with paired qid-bootstrap 95% CIs; bands are the nine-model pooled DiD. **(b)** Recipe-B seed-7 forward adapters and their matched reversals, which swap only the renderer→answer/abstain assignment. Seeds 17/23 and the parser reading are in Table A23 and Section L.

asserts a legal or eligibility conclusion the documents do not support. We find the rendering gate on publicly released RAG models evaluated on these benchmarks: holding content fixed and switching between chain and raw rendering changes how often they answer natively unanswerable items — by -10.2 percentage points for ChatQA-1.5-8B on ContractNLI (Table 1) — here raw induces *more* answers; the coupling's direction is arbitrary and benchmark-dependent (Section 4.2). There the gold label itself marks the input as admitting no supported answer, so every format-induced extra answer contradicts it: a concrete unsupported-answer channel in these released models. Analogous rendering sensitivity also appears in frontier Claude models, including claude-opus-5 and claude-sonnet-5 (Section 4.2). The gate is therefore a reliability problem for deployed RAG models on exactly the documents where an unsupported answer carries consequence.

A controlled instance, from measurement to mitigation. To study how such a gate can be learned, we train LoRA adapters on data that correlate evidence format with answer-versus-abstention labels (Section 3). With evidence content and order held fixed, enabling these adapters produces a large gate (Figure 1a). Reversing which format is paired with answering reverses the gate's direction in most models (Figure 1b), linking the behavior to the training association. To assess its effect on task performance, we vary the association's strength across seven levels. Sufficiency-oriented accuracy declines at each step as the association strengthens (Section 5.1). A decomposition of answer-or-abstain behavior reveals two accompanying trends: increasing reliance on format and decreasing sensitivity to evidence sufficiency (Section 5.3). Conversely, making evidence format independent of the training labels removes most of the gate (Section 6). Whereas prior work showed that training-time surface correlations can override prompt semantics (Shaib et al., 2025), here the same kind of correlation determines *whether a model answers at all*.

Contributions.

1. **Existence. We identify a spurious evidence-rendering gate in RAG abstention and establish its practical relevance.** In controlled experiments with format-correlated training, changing only the presentation of the same evidence, with its content and order fixed, shifts answer-or-abstain decisions in eight of nine models. We also observe format-dependent answering in released RAG models and frontier Claude models on inputs with insufficient evidence. In these cases, changing the format can increase unsupported answering (Section 4).
2. **Cause. We characterize how the gate is learned.** Reversing the training association reverses the gate, and a seven-rung dose ladder lowers sufficiency-oriented accuracy at every rung as the correlation strengthens. An audit of released post-training corpora finds model-visible surface cues that predict the abstention supervision, and injecting such an

association into a public training recipe is sufficient to produce the gate. Behavioral and internal measurements track rendering reliance rising as evidence-state sensitivity weakens, and a decision-loss intervention causally raises evidence-state sensitivity and its internal margin (Section 5).

3. **Mitigation. We mitigate the gate at training and inference time.** Balancing the renderings — marginally or by exact crossing — recovers sufficiency-oriented accuracy by up to $+0.37$ over the format-correlated recipe, and paired-rendering decision rules lower unsupported answering on frozen checkpoints relative to binary rules (Section 6).

2 Related Work

Spurious and shortcut learning. Shaib et al. (2025) is the closest behavioral account: syntactic–domain correlations absorbed during training can override prompt semantics. We study a different object, the epistemic answer-versus-abstain decision over retrieved evidence, and study it interventionally by engineering, dosing, crossing and reversing the correlation. Feature-learning, simplicity-bias, gradient-starvation, implicit-bias and group-robustness connections are in Section O.

RAG sufficiency and abstention. A growing line asks whether RAG systems answer only when the retrieved context suffices (Joren et al., 2025; Song et al., 2025; Qiu et al., 2026; Ling et al., 2025), and deployed recipes such as ChatQA (Liu et al., 2024) and OCC-RAG (Savkin et al., 2026) train for grounded answering with abstention. These works address evidence *content*; holding content byte-identical, we identify a *rendering* axis that gates the same decision, in a controlled recipe and in released models from these lines.

Format sensitivity and internal analyses. Sclar et al. (2024) quantify large task-accuracy swings under semantically equivalent prompt-format changes; here the format *gates* the epistemic answer-versus-abstain decision over retrieved evidence, and this gate is traced to the training-time renderer–target correlation through matched reversal and a graded dose response and mitigated by exact crossing (Section O develops the counterfactual-balancing connection). Ling et al. (2025) identify late-layer abstention-override artifacts; our depth profile agrees, and we add a content-fixed 2×2 design that measures internal format and evidence-state margins factorially across checkpoint trajectories and aligns them with a manipulated decision-loss weight (Section J). Checkpoint-resolved trustworthiness probing (Qian et al., 2024) is discussed in Section O.

3 Problem Formulation and Experimental Framework

3.1 Task and testbed

A RAG *reader*, the model reading the retrieved evidence, must decide whether to *answer* or *abstain*. We hold the evidence content fixed, vary only its *rendering*, and ask whether the decision tracks the rendering rather than evidence sufficiency. A content-fixed 2×2 design crosses evidence state (sufficient / insufficient) with renderer (chain / raw); within each evidence state the two renderings contain the same ordered items, so any systematic renderer dependence of the answer rate is a rendering gate by construction (Section L). The renderers serialize item i of the same ordered evidence as

chain `Step {i}.  [{evidence_id}] {title} sentence {sent_id}:  {text}`
raw `{title}:  {text}`

joined by newlines and blank lines; only the scaffold tokens differ, so chain carries more tokens (controlled in Section 4.3).

Evidence-conditioned policy. Decisions are scored against the provided evidence under the complete-groundedness convention of Section 1: on an insufficient-evidence input every substantive answer is a policy error, whether it matches the withheld gold (factually right but evidence-unsupported) or not (additionally reference-incorrect under the scoring rule); *unsupported* concerns an answer's relation to the evidence, not its truth.

3.2 DATA, READERS, AND TRAINING CONDITIONS

Data and the audited panels. Training data derive from HotpotQA (Yang et al., 2018) on a frozen 2,000-question split: a row pairs the question either with its retrieved evidence chain (answer target) or with gold-free same-question distractor paragraphs (abstention target; construction and emission rules in Section L). Evaluation panels derive from 2WikiMultiHopQA (Ho et al., 2020) with *constructed insufficiency*: a deterministic selector yields a sufficient variant and an insufficient variant in which all required support is removed and any item containing the gold answer or its aliases is excluded; a disjoint MuSiQue (Trivedi et al., 2022) panel is built the same way and held out as a *test* panel throughout. Because the selector yields only *candidate*-insufficient inputs, independent external LLMs audit every one for answerability: removing the thirteen cases identified as answerable (four 2Wiki, nine MuSiQue) leaves audited panels of 750 2Wiki and 2,382 MuSiQue sets (the adversarial residual-answerability stress analysis is in Section N). Primary analyses use development splits; every held-out panel above is test data, used for evaluation only and never for tuning, calibration or selection (the full accounting is in Section M).

Readers. A *gate-breadth fleet* of nine 0.5–7B instruction- or chat-tuned checkpoints from six families, each carrying the recipe's LoRA adapters (Hu et al., 2022), tests whether the gate appears across families and scales. A *core intervention set* of base checkpoints — Llama-3.2-3B and Qwen3-8B-Base (hereafter Qwen3-8B), extended to OLMo-2-7B — carries the dose, balancing and decision-loss analyses, which need repeated training under matched data and objective. Released readers (ChatQA, TrustAlign, OCC-RAG) and two frontier assistant systems (claude-opus-5 and claude-sonnet-5) are read off the shelf and never retrained; a separate pair of fresh Qwen2.5-Instruct 3B/7B models, post-trained with the public TrustAlign recipe on corpora we construct, serves the causal test of Section 5.2 (rosters in Section L).

Training arms. A training recipe sets the renderer–target correlation through a dose $g \in [0, 1]$: each row's renderer matches the target with probability $(1+g)/2$, so $g{=}1$ pairs chain with answer and raw with abstain (the fully correlated arm A0), $g{=}0$ makes the renderer a per-row fair coin (the marginal arm MARG), and intermediate arms ($g{=}0.5, \ldots, 0.95$) interpolate; each arm realizes a measured counterexample share $\hat{\rho}$ — the fraction of training rows whose renderer opposes the correlation — from 0% (A0), through 2.52%–24.01% across the intermediate doses ($g{=}0.95, \ldots, 0.5$), to 49.37% (MARG). The *exact crossing* arm (CROSS-CE) instead pairs every item with both renderings, and the *matched reversal* swaps only the renderer→target assignment (Table A22, Section L).

3.3 MEASUREMENTS

Gate. On a 250-question 2Wiki development panel the statistic is the per-reader *adapter answer-rate DiD*, the adapter-induced change in the chain-minus-raw answer rate within an evidence-state stratum (*cell*: one (state, renderer) pair; *insufficient stratum*: that condition under both renderings), with the nine-reader pooled DiD as the headline. An *answer* is a response that commits to a substantive answer to the question, decided by a blinded judge that reads every response of the fleet readouts under opaque identifiers, with high second-judge agreement (98.5%; Section L). The frozen deterministic parser that scored these readouts is kept as the historical reading: it counts an output as an answer unless its parsed answer contains one of six listed abstention phrases (so empty outputs and off-list refusals count), inflating its rates mainly on the adapter-off legs (per-series values and the empty-corrected variant: Section L).

Policy accuracy and its decomposition. The core-arm statistics below (SOA, Equation (1), their multi-seed and decision-loss extensions) are parser-scored on the full panels; re-judging the seed-7 dose, loss and give-back readouts in full semantically preserves every registered trend (Table A26). On a 750-question 2Wiki development panel the training-side outcome is free-generation *sufficiency-oriented accuracy* (SOA): sufficient-cell success = answered *and* correct, insufficient-cell success = abstained, fixed denominator $4n$ over the four cells (n the number of panel questions, 750 here), so SOA scores whether the policy *acts* correctly. Writing $p(s, f) = \mathbb{P}(\text{answer} \mid s, f)$ for the answer rate at evidence state $s \in \{+1, -1\}$ ($+1$ sufficient, -1 insufficient) and renderer $f \in \{+1, -1\}$ ($+1$ chain, -1 raw), the additive probit

$$p(s, f) \;=\; \Phi(u \cdot s + v \cdot f + b) \tag{1}$$

fitted per training arm (excluding the saturated A0 arm; Φ the standard normal CDF) summarizes evidence-state sensitivity u, renderer sensitivity $v = v_{\text{policy}}$, and answer propensity b. Equation (1) parameterizes the answer/abstain *action*, not accuracy; linking (u, v, b) to SOA also needs the sufficient-evidence conditional correctness (Section G). Here u measures sensitivity to the constructed evidence state, not semantic sufficiency judgement directly. The insufficient variant also differs in item count and length, but evidence-state separations persist when these are matched (Section B). Internal analogs $\hat{u}, \hat{v}$ are the evidence-state and renderer components of a layerwise logit-lens answer margin on frozen Llama-3.2-3B checkpoints at the *decision position* (the last prompt token); linear probes separately test decodability there (Section J).

External systems. The statistic is the content-fixed chain−raw answer-rate difference on unanswerable items. Released readers on ContractNLI and ShARC emit short constrained answers scored by a deterministic parser (Section K); frontier systems emit unconstrained prose, decided by a frozen three-annotator model-adjudication protocol blinded to rendering except where a response itself reveals it (Section F). Inference rules over a frozen reader's two readings are defined in Section 6.2 and scored by *unsupported-answer rate* (substantive answers on unanswerable items) and *correct coverage* (answered-and-correct rate on sufficient items). The corpus intervention retrains the public TrustAlign recipe under two matched corpora that differ only in whether a renderer marker is correlated with the label; its estimand is the correlated-minus-decorrelated contrast in the insufficient-item chain−raw answer rate (Section 5.2).

4 Existence: A Content-Fixed Rendering Gate

This section shows that the decision can track rendering — under a controlled recipe, off the shelf, and not through added length or the trained scaffold alone — for Section 3's ordinary pair of renderings.

4.1 A content-fixed rendering gate across nine readers

In one format-correlated abstention-tuning recipe, enabling the learned LoRA adapter across nine tested 0.5–7B readers on 250 development questions produces large chain-versus-raw answer-rate interactions with ordered evidence items held identical. Deterministic training at three seeds yields pooled adapter DiDs of $+0.772$ $[0.757, 0.787]$ on sufficient and $+0.703$ $[0.682, 0.723]$ on insufficient items; eight of nine readers have positive insufficient-stratum effects with per-seed 95% intervals excluding zero, and Qwen2.5-7B is unresolved at two seeds (Figure 1a). The frozen parser yields larger pooled DiDs on the same series ($+0.802$, $+0.780$; Section L). The fleet uses a shared plain-text interface; a native-template confirmation for SmolLM2 is in Section L.

4.2 Analogous rendering sensitivity off the shelf

Rendering sensitivity recurs in off-the-shelf readers we never trained — ChatQA, TrustAlign and OCC-RAG on the natively unanswerable subsets of ContractNLI and ShARC (Section 3), where any format-induced extra answer is unsupported. All four evaluated readers shift in the same, negative direction on ContractNLI, three with intervals excluding zero, while ChatQA, the one reader answering often enough on ShARC to resolve an effect, shifts positively there (Table 1a, Section K). Which rendering promotes answering is contingent on model, task and prompt frame; a negative shift is the same failure — format, not sufficiency, moving the decision.

The same sensitivity appears in Claude. Read off the shelf (Section 3), scored by semantic adjudication rather than the parser above, claude-opus-5 answers $+12.8$pp $[+9.9, +15.8]$ more often under chain than raw on the same constructed-insufficient 2Wiki panel, in the induced gate's direction (Table 1b). For claude-sonnet-5 on that panel the interval includes zero; on ContractNLI it shifts by -2.8pp $[-4.8, -0.8]$, while claude-opus-5 resolves no shift there ($+0.2$pp $[-1.6, +2.1]$). Direction and magnitude vary by system and task. A preregistered repeated-generation check (four fresh generations per rendering, 200-question subset fixed in advance) reproduces the claude-opus-5 gate at $+13.1$pp $[+9.4, +17.3]$, positive in every replicate (channel, protocol and limitations in Section F).

Table 1: Content-fixed chain−raw answer-rate differences on n unanswerable items (cluster-bootstrap 95% CIs); negative means more answers under raw. All format-induced extra answers are unsupported. Values have one decimal; differences are computed before rounding. Panels share the estimand and sign but differ in split, output channel and scoring and are not directly comparable (Sections F and K).

System	Dataset / split	n	chain / raw (%)	chain−raw [95% CI]
(a) Released RAG weights: output-specific deterministic decision rule				
ChatQA-1.5-8B	ContractNLI dev	412	13.1 / 23.3	-10.2pp $[-14.2, -6.3]$
ChatQA-1.5-8B	ShARC dev	138	31.9 / 23.2	$+8.7$pp $[+0.7, +16.7]$
TrustAlign-Qwen2.5-3B	ContractNLI dev	412	23.1 / 38.1	-15.0pp $[-20.4, -10.2]$
TrustAlign-Qwen2.5-7B	ContractNLI dev	412	55.3 / 76.9	-21.6pp $[-26.3, -17.1]$
OCC-RAG-1.7B	ContractNLI dev	412	30.6 / 34.7	-4.1pp $[-9.3, +1.0]$
(b) Off-the-shelf frontier assistants: agent channel; three-model semantic adjudication				
claude-opus-5	2Wiki dev	740	37.4 / 24.6	$+12.8$pp $[+9.9, +15.8]$
claude-sonnet-5	2Wiki dev	740	23.0 / 21.1	$+1.9$pp $[-1.2, +5.1]$
claude-opus-5	ContractNLI test	884	37.3 / 37.1	$+0.2$pp $[-1.6, +2.1]$
claude-sonnet-5	ContractNLI test	884	25.2 / 28.1	-2.8pp $[-4.8, -0.8]$

4.3 Not length alone, and not only the trained scaffold

The scaffold gates at matched length. Padding raw to chain's exact token length under two independent filler families produced no answering on constructed-insufficient inputs ($0/750$, both readers, both fillers), while chain retained the gate at that length (0.996 Llama-3.2-3B / 0.925 Qwen3-8B): added tokens alone do not explain the gate (matched-content, verbal-scaffold and paraphrase arms in Table A1, Section A).

Generalization to unseen renderers varies by reader. Holding the evidence items fixed, we re-render the 2Wiki and MuSiQue panels under seven renderers the adapters never saw — five *structural* layouts (numbered, XML, JSON, bulleted, delimiter-separated) and two *lexical* controls that keep the chain scaffold but swap its two content words ("Step"→"Stage"/"Item", "sentence"→"statement"/"part") — and evaluate the frozen adapters without retraining. The family is parser-scored; the two primary readers' (seed-7) insufficient-evidence channel is re-judged in full semantically, re-deriving their classifications below, while OLMo-2-7B stays parser-scored (Section C). Writing D_r for an adapter's renderer-versus-raw answer-rate difference on insufficient evidence, the transfer statistic is $G_r = D_r^{\mathrm{A0}} - D_r^{\mathrm{MARG}}$ (2Wiki chain anchor $+0.95$ for Llama-3.2-3B, $+0.91$ for Qwen3-8B; for OLMo-2-7B the g=0.95 dose arm stands in for A0). Every reader transfers to a lexical control (lex-syn: Llama-3.2-3B $+0.94$; Qwen3-8B $+0.08$ on 2Wiki, $+0.47$ on MuSiQue), so the gate is not tied to the scaffold's words. Structural transfer is reader-dependent: Llama-3.2-3B carries a substantial gate to the numbered, XML and JSON layouts ($+0.19$ to $+0.79$ across the two panels), each flipped to a negative value by the matched-reversal adapter, whereas Qwen3-8B's structural G_r are an order of magnitude smaller (largest $+0.037$ $[+0.029, +0.046]$, MuSiQue XML) and OLMo-2-7B's structural contrasts are small and dataset-dependent. Exact crossing leaves only small residual format differences: the largest $|D_r^{\mathrm{CROSS}}|$ across every unseen renderer and panel of the two primary readers is 0.020 (OLMo-2-7B has no crossing arm; Table A3, Section C).

5 Cause: How the Gate Is Learned

5.1 Reversing and dosing the training correlation

A matched reversal ties the direction to the learned renderer–target correlation rather than to the evidence: under Recipe B (Section L), swapping only that assignment flips the insufficient-stratum direction for 8/9 readers (Figure 1b); Phi-3.5-mini, the one reader that does not reverse, has negative estimates in both directions (values in Section L); under Recipe A it carries the gate at every seed (Table A23).

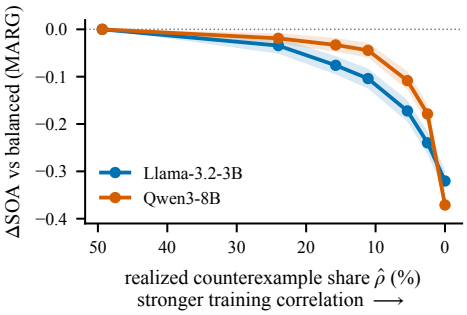

Figure 2: SOA (Section 3) falls monotonically from MARG to A0 as the realized counterexample share $\hat{\rho}$ shrinks, i.e. as the training correlation strengthens (750-question 2Wiki development panel; seed 7). Bands are simultaneous 95% max-statistic intervals; all six contrasts to MARG are negative.

On the 750-question development panel, the seven measured arms from MARG to A0 formed a monotone ladder (Figure 2, Section G): as realized counterexamples decreased from 49.37% to 0%, SOA decreased at every measured rung for both readers, and every non-baseline contrast was negative under the simultaneous max-statistic 95% bands, including the first $49.37\% \to 24.01\%$ step.

5.2 Shortcut opportunity in released corpora, and a controlled corpus intervention

The shortcut opportunity exists in released post-training data. A corpus feature audit of two released RAG post-training resources shows that model-visible surface cues predict the answer-versus-abstain supervision these readers are tuned on. In the TrustAlign SFT and DPO recipes structural features add $+3.61$ and $+3.46$ AUROC points beyond the length/count channels; most of the ChatQA mixture's separability is instead source-mediated, its visible features ranking abstaining examples under question-clustered cross-validation about as well as a *source-identity oracle* that also sees provenance metadata the model never sees. Because abstaining examples are rare, this identifies a mixture-level shortcut *opportunity*: ranking far above chance while absolute precision at the operating rate stays low (Table A6, Section D).

From association to cause: a controlled corpus intervention. The audit is observational, so we test whether a model-visible surface correlation of this kind can *cause* the gate, using an injected marker as a controlled stand-in for the audited natural features. We build two training corpora that differ in one thing. Each evidence block carries one single-token marker from a four-word family, a scaffold pole {Step, Stage} and a plain pole {Doc, Source}, so the renderings are exactly equal in length. The *correlated* corpus fixes the pole by the label — a scaffold marker on every answer row, a plain marker on every abstain row (realized marker–label AUROC 1.000) — and the *decorrelated* corpus permutes the same markers independently of the label, preserving each marker's frequency (AUROC ≈ 0.50; Section D); everything else — question text, evidence, order, labels, marker frequencies, token length — is identical.

We then run the public TrustAlign SFT+DPO recipe as fresh full-parameter post-training of Qwen2.5-Instruct 3B and 7B on each corpus, three seeds per reader. Under the correlated corpus the marker almost entirely determines answer-versus-abstain behavior: the marker-driven answer-rate difference G on insufficient evidence reaches 1.000 (3B) and 0.997 (7B), against ≤ 0.029 under the decorrelated corpus. The correlated-minus-decorrelated contrast runs $+0.970$ to $+1.000$ across both readers and both panels, all four 95% CIs above zero (Table A8). Injecting the renderer–label association is therefore *sufficient* to create the gate — at ceiling strength, where coverage also collapses without the marker. Complementary controls find rendering sensitivity in both starting models before this training and no detectable gate change under matched-dose partial redistribution of the data's natural quote and double-space associations; both concern these starting checkpoints and this recipe (Section E).

5.3 What the reader learns

The fewer the training counterexamples, the more reliable the format cue. A minimal surrogate (Section G) predicts the rising trend in v below; on a balanced panel any nonzero v costs policy accuracy, maximized at $v = 0$ for fixed $u > |b|$. The surrogate holds representations fixed, so its evidence term is constant. The decline of u below is a property of the training process itself.

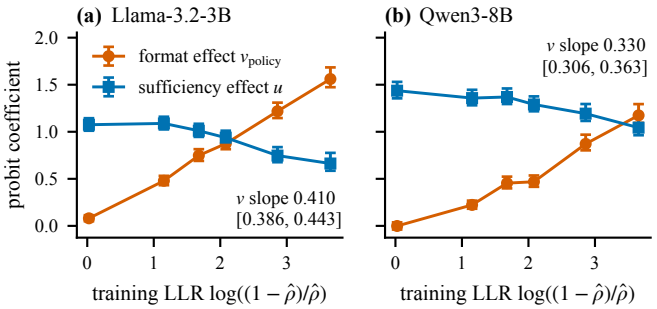

Figure 3: Policy-level probit effects versus training log-likelihood ratio (750-question 2Wiki development panel; seed 7; six nonsaturated interior arms): format v_{policy} and evidence-state u, with qid-bootstrap 95% CIs (Table A12a).

Fitting Equation (1) per dose arm, the fitted format effect v_{policy} rises while the evidence-state sensitivity u declines as the training correlation strengthens, for both readers (u as constructed in Section 3): the v_{policy}-versus-LLR slope is 0.410 $[0.386, 0.443]$ (Llama-3.2-3B) and 0.330 $[0.306, 0.363]$ (Qwen3-8B), and the decline in u is CI-separated for both (Figure 3; Table A12a). The trend is robust across link functions and a per-question hierarchical model and reproduces at seeds 17/23 and on OLMo-2-7B (Tables A11 and A14; Sections G and H).

The two tendencies also move within a single run. Holding the dose fixed and comparing the step-48 and step-249 checkpoints, all six Llama-3.2-3B configurations (three doses, two seeds) show a *give-back*: evidence-state sensitivity u recovers, by up to $+0.725$ $[0.654, 0.800]$, while format reliance v recedes, and Qwen3-8B reproduces the direction (Section H). On panels matching item count and token length across evidence states, the behavioral separations persist, u retains its overall decreasing dose trend, and the $g{=}0.8$ give-back holds for both readers (Section B). Layerwise probes and a logit lens track the internal analogs $\hat{u}, \hat{v}$ at the decision position (logit-scale margins, not commensurate with the probit coefficients); the internal trajectories share the evidence-state recovery, not the format decline. The internal format margin turns positive before the behavioral gate emerges, while the evidence-state margin dips and then recovers by $+2.035$ $[+1.948, +2.123]$ over that give-back window, ending above a format margin that shows no CI-separated decline: internally the give-back recovers the evidence-state margin *alongside* a persistent format signal. These correlational reads, with decodability at initialization, depth profiles, and dose tracking, are in Section J.

What the decision-loss intervention changes. We fork a common Llama-3.2-3B $g{=}0.8$ step-48 checkpoint and reweight the *decision-token loss* L_{dec} — the loss on the correctly labelled abstain/answer decision tokens — over a $\{0.25, \ldots, 2.0\}\times$ ladder. Increasing the L_{dec} weight raises the endpoint u monotonically at seeds 7 and 17; a complementary *remaining-token* (L_{rest}) ladder at near-equal total mean-gradient norms moves u oppositely, so the matched-gradient control attributes the increase to decision-token weighting beyond total gradient scale: D_u, the L_{dec}-minus-L_{rest} difference in the endpoint change of u, is $+0.522$ $[0.424, 0.631]$ at seed 7 and reproduces at seed 17, while its counterpart D_v varies with seed. The internal evidence-state margin tracks the same manipulation at both seeds: loss reweighting causally changes both u and its internal margin $\hat{u}$ (Table A13, Table A18a, Figure A2b). Together, u, v_{policy} and the format-effect share of Section H (Figure A3) give an account consistent with feature learning under spurious correlations (Qiu et al., 2024; Shah et al., 2020; Pezeshki et al., 2021) and motivate the inference-time rules of Section 6.

6 Mitigation at Training and Inference

6.1 Training-stage mitigation

The two balancing recipes of Section 3 — *marginal balancing* (MARG-CE) and *exact crossing* (CROSS-CE) — recover most of the accuracy lost to A0 (Figure 4), all arms using matched presentations, optimization steps and token budgets (Table A21). Exact crossing improves SOA over A0 by $+0.3453$ $[0.3297, 0.3613]$ (Llama-3.2-3B) and $+0.3747$ $[0.3603, 0.3887]$ (Qwen3-8B) and attains the highest SOA of any recipe on both readers on this panel; marginal balancing reaches essentially the same accuracy (paired CROSS$-$MARG $+0.0220$ $[0.0087, 0.0353]$ at 3B, $+0.0030$ $[-0.0060, 0.0117]$ at 8B), and across seeds and datasets the two balanced recipes' point estimates lie within ± 0.03 SOA with varying sign (Table A15).

Robustness across seeds and datasets. Both balanced recipes improve SOA over A0 for both readers at seeds $\{7, 17, 23\}$ on the 2Wiki panel and the disjoint 2,382-question MuSiQue test panel: all 24 treatment−A0 paired qid-bootstrap 95% CIs lie above zero ($+0.30$–$+0.38$ on 2Wiki, $+0.196$–$+0.264$ on MuSiQue; Table A15). On a judged subset of these arms, semantic and parser SOA differ by ≤ 0.010 (Table A25).

On 2Wiki (three seeds, semantic reading), GroupDRO on the same skewed data (g=0.8) improves on ERM (Llama-3.2-3B, Qwen3-8B) in SOA ($+0.054$, $+0.011$) and in worst-group success, the lowest success rate among the four evidence-state×renderer cells ($+0.143$, $+0.132$), yet stays below the SOA of balanced training; its SOA gain shrinks as the skew weakens, while on balanced data it raises Qwen3-8B's worst-group success ($+0.041$ $[+0.029, +0.056]$; Section I).

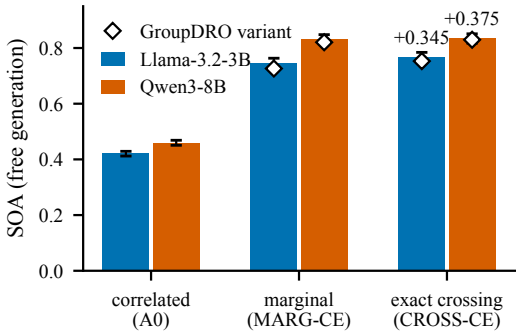

Figure 4: Training-stage mitigation on the 750-question 2Wiki development panel (seed 7). SOA with 95% CIs; labels give CROSS-CE−A0. Diamonds: GroupDRO (no SOA gain on balanced data).

6.2 Inference-stage mitigation

For off-the-shelf readers the account gives an inference-time route: read each item under both renderings and suppress what changes between them. Each rendering yields a decision margin (m_c chain, m_r raw): the teacher-forced mean per-token log-probability of the best answer label minus the abstain label's; writing $m_f = h + \delta f$ for $f = \pm 1$ separates a renderer-even score h from a renderer-odd component δ. Binary **AND** answers only when both renderings answer. **Margin fusion** answers iff $S_\lambda = \frac{1}{2}(m_c + m_r) - \lambda\,|m_c - m_r| = h - 2\lambda|\delta| > \tau$, so λ penalizes disagreement (λ=0 cancels the odd part; λ=$\frac{1}{2}$ is the maximin), and sweeping τ traces an operating curve on the (unsupported-answer rate, correct-coverage) plane. **Discordance adjudication**, for a reader that generates rather than scores labels, keeps agreed decisions and otherwise asks a format-neutral sufficiency question over the raw-rendered evidence (verbatim rule in Section K).

On ChatQA-1.5-8B, binary AND produces unsupported answers on 0.114 of unanswerable items at 0.341 correct coverage; in-sample at that coverage, margin fusion (λ=0) cuts the rate to 0.080 (-3.4pp). On held-out documents, the single-format-threshold comparison is unresolved: a threshold on the raw margin alone, chosen on calibration documents, reaches 0.083 unsupported at 0.347 correct coverage, against fusion's 0.090 at 0.347 ($+0.007$ $[-0.021, +0.033]$). OCC-RAG-1.7B's discordance adjudication cuts the unsupported-answer rate by 7.04pp versus chain-only, with similar observed sufficient-state answer rates, at a different point of that trade-off than binary AND (Table A19).

7 Discussion and Conclusion

Limitations and future directions. Future work includes direct-API access and repeated generations beyond Section F's subset; direction-level interventions testing whether Section J's internal directions carry the decision; and characterizing the renderer association left after exact crossing.

Conclusion. Evidence rendering can gate abstention in RAG, with a large learned gate; analogous sensitivity appears in released models — an unsupported-answer channel — and a deployed frontier assistant. Gate direction follows the engineered training correlation, accuracy falls throughout the dose ladder, released post-training corpora offer the same shortcut, and injecting the association into a public recipe is causally sufficient. Behavior decomposes into opposing trends in rendering reliance and evidence-state sensitivity. Balanced training recovers most lost accuracy. Inference-stage paired rules reduce unsupported answers on frozen weights relative to binary rules, but not resolvably over a calibrated single-format threshold, at about twice the scoring cost (Section K). Uncontrolled rendering can turn sufficiency questions into systematic unsupported answers.

AI USE STATEMENT

Generative AI tools assisted with writing and language polishing, and with information retrieval and discovery, including finding related work. They also served as LLM-based reference judges in the answer-matcher validation, frontier-response adjudication, the answer-commitment adjudication of Section L, and panel-answerability audit (Sections F, L and N).

REPRODUCIBILITY STATEMENT

The evaluation panels, splits and the thirteen excluded question identifiers are given in Sections M and N. The chain and raw renderers, the readout and training prompts, the decoding settings, the answer parser, the abstention predicate and both answer matchers, the per-recipe training hyperparameters with the provenance of each value (set explicitly or recorded package-version default), and the three readings of *answer* are given in Section L, with the full nine-reader per-series readings as a machine-readable table in the supplementary material; the marker family, token ids, assignment rules and full training configuration of the corpus intervention in Section D; and the frontier-lane adjudication protocol with the annotators' identities, effort settings, blinding, batching, ontology, aggregation rule and verbatim instruction in Section F. Readouts on weights we control or download use deterministic (greedy) decoding under a frozen prompt and parser; the frontier-system lane of Section 4.2 instead issues one generation for each (system, item, rendering) triple in its primary full-panel evaluation through an agent/subscription channel whose decoding we do not control (a preregistered $k{=}4$ repeated-generation supplement on a representative subset is reported in Section F). Construction, training, readout and analysis code, together with pinned environment specifications, is available in the anonymized repository.

REFERENCES

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, 2021.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, 2019.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Yuta Koreeda and Christopher D. Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Zipeng Ling, Shuliang Liu, Yuehao Tang, Junqi Yang, Shenghong Fu, Chen Huang, Kejia Huang, Yao Wan, Zhichao Hou, and Xuming Hu. LLM abstention can be a prompt artifact, in addition to genuine uncertainty. *arXiv preprint arXiv:2507.16199*, 2025.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. ChatQA: Surpassing GPT-4 on conversational QA and RAG. In *Advances in Neural Information Processing Systems*, 2024.

Meta AI. Llama 3.2 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md`, 2024.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Feature learning in the presence of spurious correlations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, 2024.

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*, 2026.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 8346–8356, 2020b.

Maksim Savkin, Mikhail Goncharov, Alexander Gambashidze, Alla Chepurova, Dmitrii Tarasov, Nikita Andriianov, Daria Pugacheva, Vasily Konovalov, Andrey Galichin, and Ivan Oseledets. OCC-RAG: Optimal cognitive core for faithful question answering. *arXiv preprint arXiv:2606.00683*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Chantal Shaib, Vinith M. Suriyakumar, Byron C. Wallace, and Marzyeh Ghassemi. Learning the wrong lessons: Syntactic-domain spurious correlations in language models. In *Advances in Neural Information Processing Systems*, 2025.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *International Conference on Learning Representations*, 2025.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, et al. 2 OLMo 2 furious (COLM's version). In *Conference on Language Modeling (COLM)*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10, 2022.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. In *Advances in Neural Information Processing Systems*, 2021.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

A MATCHED-LENGTH RENDERER CONTROL (SCAFFOLD VERSUS LENGTH)

Design. The matched-length control of Section 4.3 re-renders the 750-question development panel under thirteen matched-content conditions and reads each condition out on the frozen adapters (A0, MARG-CE, CROSS-CE; Llama-3.2-3B and Qwen3-8B; seed 7) with no retraining, so every cross-condition difference is attributable to the rendering alone. For each question and evidence state, all conditions render the identical ordered evidence items. The padded-raw controls add no scaffold: they match the chain rendering's total token length exactly under each reader's own tokenizer (zero residual tokens per item), with the padding distributed across item positions rather than concentrated at the tail, under two independent filler families (pseudo-Latin and natural English prose); a shorter padding target matches the compressed-scaffold length. The component conditions add one scaffold ingredient to raw — the item separator, the bracketed identifier, or the verbal scaffold; the verbal-scaffold condition is itself a composite of the ordinal step wording and the sentence-identifier wording (with raw's separator and no identifier bracket), so the component conditions index scaffold ingredients rather than decompose them orthogonally. The compressed-scaffold condition keeps the chain's ordinal, identifier, and sentence-number components but drops the literal words "Step" and "sentence", shortening the scaffold by changing its wording and its length together. Three paraphrase conditions reword the chain scaffold while preserving its structure and meaning. The statistic is the answer rate on constructed-insufficient inputs — under the answerability policy of Section 3, the unsupported-answer rate — with paired question-cluster bootstrap 95% CIs (B=10,000; n=750).

Findings. The chain-versus-length-matched-raw contrast equals the full confounded chain−raw difference ($+0.996$ for Llama-3.2-3B, $+0.925$ for Qwen3-8B, identical under both filler families), while adding length to raw induces no answering under either filler ($0/750$ in all four padded controls, both readers). The composite verbal scaffold recapitulates most of the effect ($+0.809$/$+0.851$ over raw); identifier and separator tokens contribute at most $+0.019$ (Llama-3.2-3B) and nothing for Qwen3-8B. Paraphrasing the scaffold attenuates the effect modestly for Llama-3.2-3B (chain−paraphrase $+0.067$ to $+0.315$) and nearly abolishes it for Qwen3-8B ($+0.873$ to $+0.912$), so the Qwen3-8B response is highly specific to the exact training wording while the Llama-3.2-3B response retains a structural component. The compressed-scaffold condition, which changes the scaffold's wording and its length together, retains most of the effect for Llama-3.2-3B (0.732) but not for Qwen3-8B (0.008); its contrast with chain is large ($+0.264$ for Llama-3.2-3B, $+0.917$ for Qwen3-8B). The padded controls isolate added length outside a scaffold, where it produces no answering; how length acts *within* a scaffold is not separated here, and with raw and every padded-raw condition at exactly zero the design leaves no headroom to measure a length effect inside unscaffolded renderings. Under the balanced MARG-CE and CROSS-CE adapters the same re-rendering yields only small chain−raw differences ($+0.040$/$+0.044$ for Llama-3.2-3B and $+0.012$/$+0.015$ for Qwen3-8B), consistent with the training-side mitigation of Section 6.1.

B COUNT- AND LENGTH-MATCHED EVIDENCE-STATE PANELS

Design. The constructed insufficient state of the main panels differs from the sufficient state not only in evidence content but also in item count and token length, so the evidence-state separations reported throughout could in principle ride on those surface statistics. This control rebuilds the insufficient state with both cues removed and re-reads the *frozen* adapters on the result (no retraining). For each retained question the sufficient state is kept unchanged, and the matched insufficient state replaces the key supporting items with natural non-answer-bearing distractor passages drawn from the same question's source pool, at the same item count (5) and per-item token-matched length under the reader tokenizer (tolerance $\pm 10\%$ for $2{,}615$ of $3{,}060$ replaced or retained items, $\pm 20\%$ for 353, $\pm 30\%$ for 92), with replacement positions spread over all five slots. Surface separability of the two states collapses by construction: item-count AUC $1.0 \rightarrow 0.5$ exactly, and token-length AUC 0.945–$0.967 \rightarrow 0.513$–0.521 (both readers, both renderers). 650 of the 750 2Wiki panel questions admit the matched construction, and every comparison below pairs the original and matched constructions on the same 650 questions; a MuSiQue panel is built the same way as a post-hoc validation set. A sham-intervention control applies the same replacement operation to *non-supporting* items of the sufficient state, leaving sufficiency intact, to separate the effect of the editing operation from the effect of removing support. Responses are scored under the blinded semantic reading (Section L) for

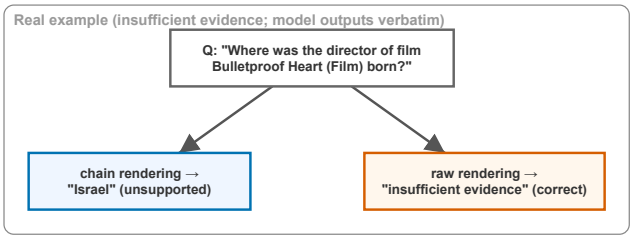

Figure A1: A real example from the 250-question development panel (Llama-3.2-3B, adapter enabled; model outputs quoted verbatim): with insufficient evidence, the chain rendering yields the unsupported answer "Israel" while the raw rendering yields the correct abstention "insufficient evidence".

Table A1: Matched-length renderer decomposition on the frozen A0 adapters: answer rate on constructed-insufficient inputs per render condition (750-question 2Wiki development panel; seed 7; paired question-cluster bootstrap 95% CIs; conditions with rate 0 have degenerate $[0, 0]$ intervals). All padded-raw conditions match the corresponding scaffolded condition's total token length exactly under each reader's tokenizer.

Render condition	Llama-3.2-3B	Qwen3-8B
raw	0.000	0.000
raw + pseudo-Latin padding (compressed-scaffold length)	0.000	0.000
raw + pseudo-Latin padding (chain length)	0.000	0.000
raw + natural-prose padding (compressed-scaffold length)	0.000	0.000
raw + natural-prose padding (chain length)	0.000	0.000
separator only	0.019 [0.009, 0.029]	0.000
identifier only	0.009 [0.003, 0.016]	0.000
verbal scaffold (composite)	0.809 [0.781, 0.836]	0.851 [0.825, 0.875]
compressed scaffold	0.732 [0.700, 0.763]	0.008 [0.003, 0.015]
chain paraphrase 1	0.848 [0.821, 0.873]	0.044 [0.031, 0.059]
chain paraphrase 2	0.929 [0.911, 0.947]	0.052 [0.036, 0.068]
chain paraphrase 3	0.681 [0.648, 0.715]	0.013 [0.005, 0.023]
chain	0.996 [0.991, 1.000]	0.925 [0.905, 0.944]

every arm, original and matched alike; the deterministic parser runs as a parallel control and agrees within 2.0 points per cell on every trained lane.

Independent sufficiency audit. A blind census in the protocol family of Section N (independent annotators, permuted evidence display, verbatim-quote verification, adjudication of disagreements) rates every retained 2Wiki matched-insufficient state insufficient (650/650) and 561 of 567 MuSiQue matched-insufficient states insufficient (the six others are excluded from all analyses), while sham-edited still-sufficient states are rated sufficient at the same rate as unedited sufficient controls (73–80% versus 79–82% verified-sufficient).

Findings. Every trained arm retains its evidence-state separation when the count and length cues are removed (Table A2): across all trained 2Wiki lanes — the dose ladder, the margin and cross-format objectives, the additional seeds, and the step-48/249 checkpoint pairs — the paired retention ratios range 1.006–1.494 with every 95% CI lower bound at or above 0.988, so no lane attenuates; on the dose-g=0.8 Llama-3.2-3B arm the fitted evidence-state sensitivity u *rises* on the matched panel by $+0.154$ $[+0.086, +0.226]$ and sufficiency-oriented accuracy rises from 0.640 to 0.683. The direction is itself diagnostic: the original insufficient contexts are the *shorter* ones, yet the readers answer them more (0.492 versus 0.331 chain-rendering answer rate on that arm), the opposite of a short-context abstention shortcut. On the Llama-3.2-3B seven-arm family (margin, five doses, cross-format CE) the simultaneous sup-norm 95% band on the matched-minus-original gap difference excludes zero for all seven arms. The trends and dynamics reported in the main text are preserved on the matched

Table A2: Evidence-state answer-rate gap (sufficient − insufficient, averaged over the chain and raw renderings) under the original and the count- and length-matched insufficient constructions, paired on the same questions (blinded semantic reading; question-cluster bootstrap 95% CIs; n=650 2Wiki, n=561 MuSiQue). The retention ratio is matched/original.

Panel	Arm (reader)	Gap orig.	Gap matched	Retention ratio
2Wiki	dose g=0.8 (Llama-3.2-3B)	0.492	0.579	1.177 [1.128, 1.228]
2Wiki	MARG-CE (Llama-3.2-3B)	0.710	0.773	1.089 [1.056, 1.124]
2Wiki	CROSS-CE (Llama-3.2-3B)	0.773	0.807	1.044 [1.017, 1.072]
2Wiki	dose g=0.8 (Qwen3-8B)	0.735	0.749	1.019 [0.998, 1.040]
2Wiki	MARG-CE (Qwen3-8B)	0.825	0.829	1.006 [0.991, 1.021]
2Wiki	CROSS-CE (Qwen3-8B)	0.787	0.801	1.018 [1.000, 1.036]
MuSiQue	MARG-CE (Llama-3.2-3B)	0.255	0.250	0.983 [0.950, 1.011]
MuSiQue	CROSS-CE (Llama-3.2-3B)	0.236	0.234	0.993 [0.967, 1.015]
MuSiQue	MARG-CE (Qwen3-8B)	0.521	0.513	0.985 [0.972, 0.995]
MuSiQue	CROSS-CE (Qwen3-8B)	0.487	0.485	0.996 [0.989, 1.004]

panel: u retains its overall decreasing dose trend (Spearman rank correlation with the dose -1.0 on every matched-panel fit under the semantic reading, and at most -0.94 on every fit under both constructions and both readings); the step-48→249 give-back holds under both constructions in all three checkpoint pairs (Llama-3.2-3B seeds 7 and 17, Qwen3-8B seed 7; u recovers and v recedes, all CIs excluding zero), with no resolvable difference between constructions in the recovery of u (all difference-in-difference CIs span zero) and one in v, a larger recession on the matched panel at Llama-3.2-3B seed 17 (-0.171 $[-0.273, -0.018]$); the matched-versus-original differences agree in sign across $3/3$ seeds in every non-saturated family, and a near-exact subset (total-length mismatch $\leq 2\%$, n=311) reproduces the headline contrasts (gap difference $+0.103$ $[+0.071, +0.135]$ on the dose-g=0.8 Llama arm). The sham control shows that the editing operation itself is not inert: across the trained lanes it lowers the answer rate on still-sufficient states by up to 0.15 under chain rendering (several Llama-3.2-3B lanes) and up to 0.17 under raw rendering (0.05 and 0.17 on the dose-g=0.8 Llama arm). Every such drop is smaller than the matched gap on the same lane, but part of any matched-minus-original increase may reflect the edit rather than the removed cues. On the MuSiQue validation panel the gaps are essentially unchanged (retention ratios 0.983–0.996; Table A2), with insufficient-state answer rates ≤ 0.02 under both constructions.

Scope. These results exclude evidence item count and token length as a sufficient explanation of the frozen readers' evidence-state separations; how far the remaining separation reflects pure semantic sufficiency judgment, and which internal mechanism carries it, remain open.

C Generalization to Unseen Renderers

Design. We ask whether the learned gate is bound to the exact training renderer or abstracts to formats the adapter never saw. Holding the ordered evidence items fixed, we render each (question, evidence-state) under nine renderers: the training anchors *raw* and *chain*; five unseen *structural* layouts — *numbered* ("1. 2. 3."), *XML* (tagged items), *JSON* (a list of objects), *bulleted*, and *delimiter*-separated; and two *lexical* controls that preserve the chain scaffold's structure but swap only its two content words — *lex-syn* ("Step"→"Stage", "sentence"→"statement") and *lex-neutral* ("Step"→"Item", "sentence"→"part"). We read the frozen adapters (A0, the matched reversal, MARG-CE, CROSS-CE) inference-only, with no retraining, so every unseen renderer is genuinely out-of-training. Responses are scored by the frozen parser, the reading under which the core arms are reported (Section L); the insufficient-evidence channel of all sixteen main-reader lanes is additionally re-judged in full under the blinded semantic reading (the semantic-re-scoring paragraph below), while the OLMo-2-7B lanes and the sufficient channel remain parser-scored. For each renderer r and adapter arm we write

$$D_r^{\mathrm{arm}} = \big[\mathrm{Ans}(r) - \mathrm{Ans}(\text{raw})\big]_{\mathrm{arm}}$$

for the within-arm renderer-versus-raw answer-rate difference on the insufficient-evidence (unsupported-answer) channel, and measure the difference-in-differences $G_r = D_r^{\mathrm{A0}} - D_r^{\mathrm{MARG}}$, the

gap under the format-correlated policy net of the balanced policy: G_r near the chain anchor means the gate transfers to r, G_r near zero means it does not. OLMo-2-7B has no A0 arm in this family; its contrast uses the g=0.95 dose arm in place of A0 (seeds 7/17/23, transfer-only, no reversal arm). The primary reading throughout this family is by effect size with its paired qid-cluster bootstrap 95% CI (B=10,000): a resolved G_r of any magnitude is reported as an effect, and no fixed threshold decides whether a transfer exists. The preregistered classification is retained as a clearly marked *registered auxiliary label*: a transfer is labelled *within-region* when its whole CI lies inside ± 0.05, and *confirmed* when G_r is positive with a CI not wholly inside that region *and* the matched-reversal adapter flips its sign. Panels are the audited 2Wiki-750 development and disjoint MuSiQue-2,382 test sets.

Reader-dependent abstraction. The gate is not restricted to the scaffold's original words: every reader shows a resolved positive G_r on at least one lexical "Step"/"sentence" swap, from $+0.076$ $[+0.052, +0.100]$ (Qwen3-8B, 2Wiki lex-syn) to $+1.000$. These controls leave the rest of the scaffold in place, so they bound reliance on those two words rather than on its tokens generally. On unseen *structural* layouts the readers diverge (Table A3). For Llama-3.2-3B the numbered, XML, and JSON layouts reproduce a substantial gate on both panels, and the matched-reversal adapter flips each to a CI-separated negative value (bulleted also carries a resolved gate on MuSiQue; the delimiter layout shows no resolved transfer on either panel). For Qwen3-8B the unseen structural G_r are small — the largest, MuSiQue XML, is $+0.037$ $[+0.029, +0.046]$, resolved but an order of magnitude below its own lexical transfer and below Llama-3.2-3B's structural transfer (every structural CI also lies within the registered auxiliary region) — so its abstraction is predominantly lexical rather than structural. OLMo-2-7B, pooled over seeds $\{7, 17, 23\}$ (g=0.95 arm versus MARG, transfer-only), shows all-seed lexical transfer on both panels (lex-syn/lex-neutral $+0.292$/$+0.170$ on 2Wiki, $+0.146$/$+0.068$ on MuSiQue), with small *dataset-dependent* structural contrasts: numbered/XML/JSON $+0.040$/$+0.072$/$+0.064$ on 2Wiki, against -0.028 to $+0.033$ on MuSiQue (per-arm D_r under both baselines — pooled estimates with per-seed sign-consistency markers — in Table A4). The learned gate is thus reader-dependent: structural and lexical for Llama-3.2-3B, predominantly lexical for Qwen3-8B, and lexical with small dataset-dependent structural contrasts for OLMo-2-7B. The chain anchors quoted in Section 4.3 are the 2Wiki values.

Mitigation and equivalence audit. The mitigation statistic is the within-arm difference itself, not G_r: under exact crossing (CROSS-CE) the largest residual $|D_r^{\mathrm{CROSS}}|$ across all 28 (renderer, reader, panel) cells is 0.020 — small but partly resolved, never absent — and every cell also carries the registered within-region label; under marginal balancing the two largest residuals are Llama-3.2-3B's 2Wiki XML and JSON D_r^{MARG}, $+0.060$ $[+0.037, +0.083]$ and $+0.037$ $[+0.016, +0.059]$ (the remaining 26 cells carry the within-region label). A blind renderer-equivalence check — an independent rater scoring 216 label-randomized renderings (12 questions $\times$ 2 evidence states $\times$ 9 formats) — found every group content-fixed and answerability-equivalent in the sample, with no renderer injecting an answer, abstention, or continue-reasoning instruction; that the readers still gate on the neutral lexical control indicates the learned cue is more abstract than the human-perceptible reasoning-chain wording.

Semantic re-scoring of this family. The insufficient-evidence channel of all sixteen main-reader lanes — both readers, all four arms, both panels, all nine renderers (225,504 responses) — was re-judged in full by the blinded semantic reading of Section L, and the transfer estimator re-run with only the answer label changed. Every classification above re-derives identically: Llama-3.2-3B's structural transfer and reversal flips (2Wiki XML $+0.305 \rightarrow +0.296$, JSON $+0.437 \rightarrow +0.411$; MuSiQue XML $+0.785 \rightarrow +0.764$, JSON $+0.607 \rightarrow +0.589$), Qwen3-8B's small structural contrasts (largest, MuSiQue XML, $+0.037$ under both readings), the largest residual $|D_r^{\mathrm{CROSS}}|$ of 0.020 under both readings, and the same two 2Wiki D_r^{MARG} residuals (registered-region counts 28/28 and 26/28 unchanged). The parser and the judge disagree on 1.0% of these responses (2,267 of 225,504), all but one a response the parser scores as an answer and the judge reads as an abstention or an invalid output — the direction the structured-refusal concern predicts — and the largest per-cell coefficient changes sit on the seen anchors (≤ 0.095), not the unseen renderers. Within-run consistency of the freshly judged pairs is $286/300 = 0.953$ $[0.923, 0.972]$.

Table A3: Unseen-renderer generalization on the insufficient-evidence channel: transfer statistic $G_r = D_r^{\mathrm{A0}} - D_r^{\mathrm{MARG}}$ (frozen adapters, inference-only, parser-scored; seed 7; 2Wiki n=750 and MuSiQue n=2,382; paired qid-cluster bootstrap 95% CIs in brackets). *raw*/*chain* are the training anchors; the five structural and two lexical renderers are unseen. G_r is a difference of two rate differences, so its algebraic range is $[-2, 2]$ and interval endpoints may lie outside $[-1, 1]$. For Llama-3.2-3B the structural gate abstracts (with a matched-reversal sign flip); Qwen3-8B's structural G_r are an order of magnitude smaller (largest $+0.037$), its lexical transfer resolved on both panels for lex-syn (lex-neutral on MuSiQue).

Renderer	Panel	Llama-3.2-3B	Qwen3-8B
chain (anchor)	2Wiki	$+0.952$ $[+0.931, +0.973]$	$+0.913$ $[+0.891, +0.935]$
	MuSiQue	$+0.998$ $[+0.994, +1.002]$	$+0.886$ $[+0.873, +0.899]$
numbered	2Wiki	$+0.188$ $[+0.156, +0.221]$	-0.007 $[-0.017, +0.003]$
	MuSiQue	$+0.422$ $[+0.403, +0.443]$	$+0.009$ $[+0.004, +0.014]$
XML	2Wiki	$+0.305$ $[+0.265, +0.347]$	$+0.005$ $[-0.005, +0.016]$
	MuSiQue	$+0.785$ $[+0.769, +0.802]$	$+0.037$ $[+0.029, +0.046]$
JSON	2Wiki	$+0.437$ $[+0.396, +0.477]$	$+0.003$ $[-0.009, +0.015]$
	MuSiQue	$+0.607$ $[+0.587, +0.627]$	$+0.016$ $[+0.010, +0.022]$
bulleted	2Wiki	$+0.060$ $[+0.036, +0.085]$	-0.007 $[-0.017, +0.004]$
	MuSiQue	$+0.180$ $[+0.165, +0.197]$	$+0.005$ $[+0.000, +0.009]$
delimiter	2Wiki	$+0.000$ $[-0.015, +0.013]$	-0.004 $[-0.011, +0.003]$
	MuSiQue	-0.003 $[-0.005, +0.000]$	-0.001 $[-0.003, +0.001]$
lex-syn	2Wiki	$+0.941$ $[+0.916, +0.967]$	$+0.076$ $[+0.052, +0.100]$
	MuSiQue	$+1.000$ $[+0.997, +1.004]$	$+0.473$ $[+0.453, +0.493]$
lex-neutral	2Wiki	$+0.775$ $[+0.740, +0.809]$	-0.009 $[-0.023, +0.004]$
	MuSiQue	$+0.995$ $[+0.990, +1.000]$	$+0.126$ $[+0.113, +0.140]$

Table A4: OLMo-2-7B unseen-renderer companion statistics on the insufficient-evidence channel: the within-arm renderer-versus-raw answer-rate difference D_r under the g=0.95 dose arm and under MARG, and the transfer statistic $G_r = D_r^{g=0.95} - D_r^{\mathrm{MARG}}$, pooled over seeds $\{7, 17, 23\}$ by per-question seed-mean (parser reading; paired qid-cluster bootstrap 95% CIs; * interval excludes zero; † every per-seed interval excludes zero with a common sign). The g=0.95 arm carries $\approx 2.5\%$ training counterexamples, so this baseline is not the A0 baseline of Table A3 and the two contrasts are not directly comparable.

Renderer	Panel	$D_r^{g=0.95}$	D_r^{MARG}	G_r
chain (anchor)	2Wiki	$+0.476$ $[+0.442, +0.508]^*$	$+0.006$ $[-0.005, +0.016]$	$+0.470$ $[+0.435, +0.503]^{*\dagger}$
	MuSiQue	$+0.327$ $[+0.309, +0.345]^*$	$+0.002$ $[-0.002, +0.005]$	$+0.326$ $[+0.308, +0.344]^{*\dagger}$
numbered	2Wiki	$+0.037$ $[+0.025, +0.049]^*$	-0.003 $[-0.009, +0.003]$	$+0.040$ $[+0.027, +0.053]^{*\dagger}$
	MuSiQue	$+0.015$ $[+0.011, +0.020]^*$	$+0.002$ $[-0.001, +0.005]$	$+0.014$ $[+0.009, +0.019]^{*\dagger}$
XML	2Wiki	$+0.076$ $[+0.059, +0.093]^*$	$+0.004$ $[-0.006, +0.013]$	$+0.072$ $[+0.054, +0.090]^{*\dagger}$
	MuSiQue	$+0.021$ $[+0.016, +0.027]^*$	$+0.004$ $[+0.000, +0.007]^*$	$+0.018$ $[+0.012, +0.024]^{*\dagger}$
JSON	2Wiki	$+0.082$ $[+0.064, +0.101]^*$	$+0.019$ $[+0.008, +0.029]^*$	$+0.064$ $[+0.045, +0.082]^{*\dagger}$
	MuSiQue	$+0.040$ $[+0.033, +0.047]^*$	$+0.007$ $[+0.003, +0.011]^*$	$+0.033$ $[+0.026, +0.041]^{*\dagger}$
bulleted	2Wiki	$+0.032$ $[+0.023, +0.043]^*$	-0.003 $[-0.010, +0.004]$	$+0.035$ $[+0.023, +0.048]^{*\dagger}$
	MuSiQue	$+0.008$ $[+0.005, +0.011]^*$	-0.000 $[-0.003, +0.002]$	$+0.008$ $[+0.004, +0.011]^{*\dagger}$
delimiter	2Wiki	$+0.018$ $[+0.011, +0.027]^*$	$+0.002$ $[-0.004, +0.008]$	$+0.016$ $[+0.007, +0.026]^*$
	MuSiQue	$+0.008$ $[+0.005, +0.012]^*$	$+0.037$ $[+0.030, +0.043]^*$	-0.028 $[-0.035, -0.022]^{*\dagger}$
lex-syn	2Wiki	$+0.289$ $[+0.260, +0.320]^*$	-0.002 $[-0.012, +0.008]$	$+0.292$ $[+0.260, +0.324]^{*\dagger}$
	MuSiQue	$+0.153$ $[+0.139, +0.166]^*$	$+0.006$ $[+0.003, +0.010]^*$	$+0.146$ $[+0.133, +0.160]^{*\dagger}$
lex-neutral	2Wiki	$+0.159$ $[+0.135, +0.183]^*$	-0.011 $[-0.020, -0.002]^*$	$+0.170$ $[+0.144, +0.196]^{*\dagger}$
	MuSiQue	$+0.069$ $[+0.060, +0.078]^*$	$+0.001$ $[-0.002, +0.004]$	$+0.068$ $[+0.059, +0.078]^{*\dagger}$

Scope. These frozen-policy results measure answer-versus-abstain behaviour on the insufficient-evidence channel; the sufficient-evidence channel differs for Qwen3-8B and is parser-scored, as are the OLMo-2-7B lanes.

D External Format Sensitivity and Operating Profiles

ChatQA-1.5-8B's held-out test estimates (evaluation only; Section 3) are -11.7pp $[-14.8, -8.8]$ on ContractNLI and $+9.4$pp $[+5.6, +13.3]$ on ShARC. Two further observations characterize the released external checkpoints. *Format sensitivity:* across the seven external checkpoints we evaluated for content-fixed format sensitivity on the 750-question panel, four show a CI-separated insufficient-item answer-rate shift (chain$-$raw; Table A5) and three have CIs that include zero; the three largest shifts — ChatQA-1.5-8B, Command-R7B, and OCC-RAG-1.7B — are negative, meaning these readers abstain *more* under the chain rendering, the opposite direction to the gate our recipe induces. *Operating profiles:* the checkpoints occupy different failure regimes. ARENA-Qwen-7B answers almost everything (sufficient-cell correctness 0.841–0.844, but 0.899–0.915 of insufficient cells answered), while the TrustAlign checkpoints over-abstain (answering 0.053–0.332 of sufficient and 0.003–0.048 of insufficient cells); because SOA rewards the conjunction of correct answering under sufficiency and abstention under insufficiency, neither profile scores well.

Corpus feature audit protocol. The audit of Section 5.2 predicts each training example's answer-versus-abstain supervision label (abstention as the positive class) from input-observable features of the example as the model sees it. Three feature families enter: the *visible instruction family* read from the prompt surface; *length/count* statistics (character and token lengths, document count); and *structural markup* (title, numbering, bullet, newline, sentence, and digit-fraction statistics). Nested logistic models add one family at a time, and the *source-identity oracle* augments the visible features with the example's source-subset label — dataset-provenance metadata invisible to the model at training time — as a reference for source-mediated separability. All models are scored by question-clustered 5-fold cross-validated AUROC — and, because abstention is rare, also by the absolute AUPRC and the precision at a fixed recall (Table A6a) — and both absolute metrics and the nested increments (ΔAUROC) carry question-cluster bootstrap 95% CIs. The ChatQA ladder uses a 29,999-row unique-row sample of the nine-subset RAG mixture (20,639 question clusters; 103 abstain rows); the TrustAlign recipes contain 8,877 rows (8,579 clusters, SFT) and 15,598 rows (3,249 clusters, DPO), where the structural increment is measured beyond the length/count channels within each recipe. Table A6 summarizes the headline comparison of Section 5.2. The full nested ladder on the ChatQA mixture runs, in cross-validated AUROC under the primary unique-row weighting: intercept 0.4377; instruction family alone 0.9324; length/count 0.9461; length/count + structural 0.9648; all visible families 0.9761; with the model-invisible source label 0.9763 — an exposure-weighted sensitivity re-analysis preserves the ordering. The structural increment beyond length/count is $+0.0186$ $[+0.0117, +0.0266]$, and the all-visible surface sits -0.0002 $[-0.0006, +0.0001]$ from the source oracle; the full ladder with every nested increment and interval is tabulated in the supplementary material.

Operational precision under rare positives. Abstention is the positive class and is rare — 103 of the 29,999 analyzed unique-row examples (0.34%); the full unweighted nine-subset corpus contains 1,202 abstentions among 314,260 rows (0.38%), and under official blend weighting over the included RAG subsets the prevalence is 0.996% — so a high AUROC reflects rank separability and can coexist with low absolute precision. Alongside the ladder AUROC we therefore report, for the models named in Table A6a, the absolute AUPRC and the precision at a fixed 50% recall, each with a question-cluster bootstrap 95% CI. For the full model-visible surface ($\mathrm{M_{ALL}}$) the AUPRC is 0.085 $[0.067, 0.113]$ and the precision at 50% recall is 0.116 $[0.092, 0.145]$: recovering half the abstaining examples with a threshold rule would admit roughly eight false positives per true one. Removing the visible instruction family ($\mathrm{M_{LS}}$) barely changes the AUROC ($0.976 \rightarrow 0.965$) but reduces the precision at 50% recall from 0.116 to 0.065 $[0.048, 0.098]$ — a 43.7% reduction. The instruction family is a proxy for the sole subset carrying the abstention labels (the per-source prevalence below), not for individual abstaining rows.

Leave-one-source-out is not identifiable. The abstention labels are not spread across the nine ChatQA subsets: they are confined to one. Only `synthetic_convqa` carries any abstaining examples (1,202 of 38,689; 3.1%), and the other eight subsets are answer-only (per-source row and abstain counts in the supplementary material). A leave-one-source-out evaluation is therefore undefined: holding out `synthetic_convqa` leaves the training folds with no positives (nothing to learn), and holding out any other subset leaves the test fold with no positives (AUROC and AUPRC

Table A5: Content-fixed insufficient-item answer-rate shift (chain−raw) for the external checkpoints with a CI-separated effect on the 750-question 2Wiki panel (cluster-bootstrap 95% CIs). Negative values mean the checkpoint abstains more under the chain rendering — the opposite direction to the gate our recipe induces. For the remaining checkpoints (ARENA-Qwen-7B, TrustAlign-Llama3.2-3B, TrustAlign-Llama3-8B) the intervals include zero.

Checkpoint	chain−raw [95% CI]
ChatQA-1.5-8B	-0.161 $[-0.188, -0.135]$
Command-R7B	-0.084 $[-0.113, -0.055]$
OCC-RAG-1.7B	-0.049 $[-0.076, -0.024]$
TrustAlign-Llama2-7B	$+0.013$ $[+0.005, +0.023]$

Table A6: Corpus feature audit of released RAG post-training resources: how well input-observable surface features of a training example predict its answer-versus-abstain supervision (question-clustered cross-validation; protocol in Section D). **(a)** ChatQA-1.5 RAG mixture (nine subsets): rank separability (AUROC) shown against the operational cost under the rare abstention rate (AUPRC and precision at 50% recall), for the model-visible surface, the same surface without the instruction family, a within-source model that additionally holds source and instruction fixed, and the model-invisible source-identity oracle (diagnostic ceiling); each panel-(a) cell is a point estimate with a question-cluster bootstrap 95% CI. Abstention is rare in the mixture (≈0.34%) but less rare within the single label-bearing source (≈3%), so AUPRC and precision compare only within a base rate; across base rates the AUROC (and the within-source increment) is the comparable quantity. **(b)** TrustAlign recipes: AUROC added by the length/count channels and by structural features beyond them (question-cluster bootstrap 95% CIs).

(a) ChatQA-1.5 mixture: rank separability vs. operational precision

Population / feature model	AUROC	AUPRC	Prec. @ 50% recall
Mixture, all visible features	0.976 $[0.971, 0.981]$	0.085 $[0.067, 0.113]$	0.116 $[0.092, 0.145]$
Mixture, no instruction family	0.965 $[0.957, 0.972]$	0.068 $[0.051, 0.096]$	0.065 $[0.048, 0.098]$
Within-source (src+instr. fixed)	0.831 $[0.816, 0.846]$	0.172 $[0.155, 0.193]$	0.194 $[0.175, 0.213]$
Source-identity oracle (diagnostic)	0.976 $[0.971, 0.981]$	0.088 $[0.068, 0.117]$	0.118 $[0.094, 0.147]$

(b) TrustAlign recipes (ΔAUROC)

Recipe	length/count	structural beyond length/count [95% CI]
SFT	$+0.0452$	$+0.0361$ $[+0.0224, +0.0488]$
DPO	$+0.0326$	$+0.0346$ $[+0.0216, +0.0474]$

undefined). Label support is bound to source, so cross-source generalization of the abstention predictor is not estimable from these resources: the pooled-mixture separability identifies a mixture-level shortcut *opportunity* within these corpora, and is silent about whether generic rendering structure predicts abstention across sources.

Within-source analysis. Because `synthetic_convqa` is the only subset with label variation, we re-run the ladder within it. There the source and the instruction family are constant by construction, so the visible-instruction model collapses onto the intercept and the all-visible and source-oracle models collapse onto the length+structural model; the informative comparison is $\mathrm{M_L}$ (length/count) versus $\mathrm{M_{LS}}$ (length/count+structural). With source and instruction held fixed the AUROC falls to 0.821 $[0.807, 0.835]$ ($\mathrm{M_L}$) and 0.831 $[0.816, 0.846]$ ($\mathrm{M_{LS}}$) — well below the pooled-mixture 0.976, confirming that most of the mixture separability was source-mediated — and the structural increment beyond length/count is small, with both intervals excluding zero (ΔAUROC $+0.0101$ $[+0.0059, +0.0142]$; ΔAUPRC $+0.0585$ $[+0.0474, +0.0719]$). The within-source positive rate (≈3%) exceeds the mixture's, so its AUPRC and precision (Table A6a) are not directly comparable across the two base rates; the comparable quantities are the AUROC and this structural increment.

Controlled corpus-feature intervention protocol. The audit above is observational; to test whether the renderer channel *causes* the content-fixed gate we intervene on the training corpus directly. We build two matched corpora that differ only in the joint assignment of an isometric single-token renderer marker to the answer-versus-abstain supervision label: a *correlated* corpus in which the marker predicts the label, and a *decorrelated* corpus in which the marker is independent of the label while the same marker vocabulary keeps the same marginal frequency. The markers are single tokens on the chain-versus-raw axis (a scaffold-pole marker versus a plain-pole marker), so the two renderings are exactly equal in length and the intervention holds evidence content, document order, supervision labels, token length, and renderer diversity fixed — only the marker–label correlation moves. On each corpus we run fresh full-parameter post-training with the public TrustAlign SFT+DPO recipe, initialized from the released Qwen2.5-Instruct 3B and 7B weights (no pretraining from random initialization), at seeds $\{7, 17, 23\}$; we then read each resulting checkpoint on the 750-question 2Wiki development panel and the $2{,}382$-question MuSiQue test panel with chain and raw markers whose content is byte-identical up to the marker word. The gate is the insufficient-item answer rate under the chain marker minus that under the raw marker, seed-averaged per question, and the reported contrast is $\Delta G = G_{\text{correlated}} - G_{\text{decorrelated}}$ with a paired qid-cluster bootstrap 95% CI (Table A8). The 3B study additionally trains the recipe's own native arm as a reference point; the 7B study is a reduced replication running only the two intervention arms, so its native answer rates speak to functioning, non-collapsed behavior rather than to preservation relative to the native recipe. The intervention establishes causal sufficiency of the renderer-marker channel at ceiling injection strength on both reader sizes; a graded dose-response, other model families, and the origin of the released checkpoints' own sensitivity are separate questions — Section E takes up the recipe's own natural typographic associations and the starting checkpoints directly. Table A7 gives the recipe's training configuration as run.

Marker family and assignment. The family has four single-token marker words in two poles — scaffold pole {Step, Stage}, plain pole {Doc, Source}. Every document line reads `{word} {idx}. [{idx}] {title}: {text}`, lines joined by newlines, with the Instruction, Question and trailing `Answer:` byte-identical to the native prompt, which only the original arm uses; readout contrasts Step (chain) with Doc (raw). Under the Qwen2.5 tokenizer the four words are single tokens (ids 8304, 19398, 9550, 3608) both line-initially and after a newline. The correlated corpus assigns the pole by label (answer → scaffold, abstain → plain) and the word within the pole by a fixed per-row hash; the decorrelated corpus takes that template multiset and permutes it within each partition (train and held-out separately, each kept in the original row order of the recipe file) in a label-independent key-hash order, so per-partition marker frequencies are identical (SFT train: Step 2,318, Stage 2,284, Doc 2,044, Source 1,994) and the assignment is independent of the label. The realized marker→label rank AUROC is 1.000 (correlated) versus 0.504 (decorrelated) on SFT and 1.000 versus 0.501 on DPO. Marker renders are token-isometric and padded to the per-example maximum with residual 0. The assignment is fully determined by the following data-construction rules (applied separately to the SFT and the DPO split of the public recipe data; the target labelled is the assistant message for SFT and the chosen response for DPO; a row is *abstain* iff its canonicalized target — casefolded, curly apostrophes straightened, "can't"/"couldn't" contracted, non-alphanumerics replaced by spaces, whitespace collapsed — contains "couldnt find an answer to your question in the search results"):

```
key(row)      = SHA256_hex(original user prompt string)
h(parts...)   = int(SHA256_hex(join(parts, U+241F)), 16)
CORR(row)     = pool[ h("corr", key) mod 2 ],   pool = (Step, Stage) if answer
                                                       (Doc, Source)  if abstain
partition     : the n_test rows with the largest keys are held out (membership only);
                n_test = r + round((256 - r)/b) * b  with r = n mod b, b = effective batch
                (b = 72 SFT -> 237 held out of 8,877; b = 64 DPO -> 238 of 15,598);
                inside each partition the rows KEEP the original row order of the public
                recipe file (the split's train.json; every row parses, none is dropped)
BALANCED      : within each partition P = (p_0, ..., p_{m-1}) in that original order,
                perm = argsort over j = 0..m-1 of h("bal", "20260821", key(p_j));
                row p_{perm[k]} receives CORR(p_k), k = 0..m-1   (same multiset, label-blind)
readout       : chain = Step, raw = Doc
```

Table A7: The public Trust-Align post-training recipe as run for the controlled corpus intervention of Section D: its released training scripts at the ICLR 2025 release revision, executed unmodified through `accelerate`, with the arms differing only in which corpus they read and in the seed. Provenance: E, set explicitly by the driver; D, the recipe's own configured value carried through unchanged. The step budgets follow from the row counts and the effective batch, and the held-out sizes are chosen so the training count divides exactly by the effective batch. Three protocol facts accompany the table: no intermediate checkpoint is saved and no convergence-based pick is made, so each stage trains its fixed step budget and its final model is used; each arm's DPO stage starts from that same arm and seed's SFT checkpoint; and while these settings are identical across arms, the execution scheduling (per-device batch size $\times$ gradient accumulation $\times$ device count, whose product is held at the effective batch above) may differ between runs.

	SFT stage	DPO stage
rows: train / held out	8,640 / 237 (E)	15,360 / 238 (E)
epochs; optimization steps	2; 240 (E)	2; 480 (E)
effective batch	72 (D)	64 (D)
learning rate	2×10^{-5} (D)	5×10^{-7} (D)
schedule; warm-up ratio	cosine; 0.1 (D)	cosine; 0.1 (D)
optimizer	adamw_torch (D)	paged_adamw_8bit (D)
DPO β	—	0.5 (D)
sequence length	2,048 (D)	2,048; prompt 1,600 (D)
gradient clip; length grouping	1.0; true (D)	— (D)
precision; attention	bf16; SDPA (E)	bf16; SDPA (E)

Table A8: Controlled corpus-feature intervention: correlating an isometric single-token renderer marker with the answer-versus-abstain supervision label — versus decorrelating it, with content, document order, supervision labels, token length, and renderer diversity all held fixed — is causally sufficient to produce the content-fixed chain-versus-raw gate. The public TrustAlign SFT+DPO recipe is run as fresh full-parameter post-training on each reader under both corpora (seeds $\{7, 17, 23\}$; protocol in Section D). **(a)** Correlated-minus-decorrelated gate contrast ΔG (paired qid-cluster bootstrap 95% CIs); all four intervals lie above zero and are per-seed consistent (ΔG is a difference of two rate differences, algebraic range $[-2, 2]$, so interval endpoints may exceed 1). **(b)** Arm behavior per reader, seed-averaged: the gate G (insufficient-item chain-minus-raw answer rate), correct coverage under each rendering, and native answer rate on ASQA / QAMPARI / ELI5. The correlated arm is renderer-keyed at ceiling on both readers (raw-render coverage 0.000: it abstains under the plain marker even on sufficient evidence). The decorrelated arm is substantially less renderer-sensitive, and near-invariant at 7B: the chain$-$raw coverage gap is 0.018 / 0.010 at 3B versus 0.005 / 0.003 at 7B, and a small residual 2Wiki gate remains at 3B. The 3B study also trains the recipe's own native arm, whose native rates (0.757 / 0.448 / 0.279) the decorrelated arm matches; the reduced 7B study omits that arm, so its native rates establish functioning, non-collapsed answering rather than preservation relative to the native recipe.

(a) Correlated $-$ decorrelated gate contrast ΔG

Reader	2Wiki (strict-750)	MuSiQue (strict-2382)
Qwen2.5-Instruct 3B	$+0.970$ $[0.959, 0.981]$	$+1.000$ $[0.998, 1.003]$
Qwen2.5-Instruct 7B	$+0.996$ $[0.993, 0.999]$	$+1.000$ $[0.999, 1.000]$

(b) Arm behavior, by reader

		Gate G	Coverage, chain / raw		Native answer rate
Corpus arm	Reader	2Wiki / MuSiQue	2Wiki	MuSiQue	ASQA / QAMPARI / ELI5
Correlated	3B	1.000 / 1.000	0.793 / 0.000	0.485 / 0.000	0.020 / 0.005 / 0.000
	7B	0.997 / 1.000	0.782 / 0.000	0.516 / 0.000	0.000 / 0.000 / 0.000
Decorrelated	3B	0.029 / 0.000	0.605 / 0.587	0.107 / 0.096	0.756 / 0.446 / 0.280
	7B	0.001 / 0.000	0.277 / 0.272	0.076 / 0.072	0.637 / 0.310 / 0.191

E NATURAL-FEATURE INTERVENTIONS AND TRAINING-STAGE CONTROLS

The intervention of Section D injects an artificial renderer–label association at ceiling strength. This appendix reports the complementary controls on the same public recipe: two matched-dose redistribution studies of the corpus's own weak *natural* typographic associations (quotes alone; then quotes $\times$ double spaces as a factorial), and a reading of the starting checkpoints that places the chain/raw scaffold gate on the training trajectory. The single-factor study trains Qwen2.5-Instruct 3B and 7B; the factorial trains the 3B reader only; the starting-checkpoint anchors cover both readers. All gate-panel readouts share the 750-question 2Wiki development panel and the blinded semantic reading (Section L) with the recipe's refusal parser as a parallel control; native answering is scored by the recipe's refusal parser and native recall by deterministic gold matching, with the semantic instrument auditing stratified native samples; the single-factor study additionally validates on MuSiQue.

Natural quote redistribution (single factor). In the public recipe's SFT data, curly-versus-straight quote glyphs in the retrieved documents carry a weak natural association with the abstain label (beyond-length probe ΔAUROC $+0.035$ $[0.024, 0.046]$; document text only — instruction and question bytes are never edited). A semi-synthetic redistribution moves this feature between examples at byte-matched edit dose and exposure (156 normalize $+$ 156 stylize rows per arm of $8{,}640$; $1{,}867$ characters each direction; residual exposure identical to the natural data), yielding arms that differ only in the correlation — preserved (98% of natural), halved (48%), and minimized (37%), i.e. a removal of 62.9% of the probe-measured association, never full removal — with the preference-stage data identical across arms. The redistribution produces no detectable change in either format-gate axis — the quote-style gate (styled versus plain quote rendering) or the chain/raw scaffold gate — on Qwen2.5-3B (4 arms $\times$ 3 seeds) or Qwen2.5-7B (preserved versus minimized $\times$ 3 seeds), at the SFT or DPO endpoint, on the 2Wiki development panel or the $2{,}382$-question MuSiQue validation panel: all nine preregistered preserved-minus-minimized contrasts are null with 95% CI half-widths of 0.13–0.9pp (e.g. 3B style/DPO -0.002 $[-0.007, +0.002]$; 7B style/DPO $+0.000$ $[-0.001, +0.002]$). The gate change under this partial redistribution is bounded at roughly one percentage point or less in eight of the nine contrasts; the widest — the 3B scaffold gate at the DPO endpoint, $+0.53$pp $[-0.40, +1.42]$pp — permits up to about 1.4pp and neither confirms nor excludes a small decrease. The trained models express no quote-style gate in any arm, including the untouched natural arm ($|G| \leq 0.004$ everywhere), so the null is measured against absent quote-style gates rather than as a reduction of an existing one; the recipe's own chain/raw scaffold gate ($\approx +0.07$ on 2Wiki) is present in every arm for the 3B reader (the 7B endpoint gates are near zero; see the training-stage paragraphs below) with no detectable difference under the tested partial redistribution — in sharp contrast to the injected association's $\approx +1.0$ on the same pipeline (Section D).

Joint quote $\times$ double-space factorial. A second, independent typographic channel is carried by the question spans: double spaces associate with the abstain label (beyond-length probe ΔAUROC $+0.031$ $[+0.024, +0.037]$; 10.6% of abstain-labelled versus 2.6% of answer-labelled questions). Four training arms of the 3B reader form a 2×2 over the two channels — preserve both / remove quotes / remove spaces / remove both — at byte-matched edit dose and exposure per feature, with the quote factor reproduced exactly from the single-factor study's audited assignment; the probe-measured removals are 69% (quotes), $\sim$$100\%$ (double spaces), and 85% (joint), three seeds per cell at the SFT endpoint. Estimands: the quote main effect is the mean over the two quote-removed cells minus the mean over the two quote-preserved cells (the double-space main effect is analogous), and the interaction is the deviation from additivity, (both-removed $-$ spaces-removed) $-$ (quotes-removed $-$ control); positive values mean a larger metric under removal; per-question values are seed-averaged within arm first, and CIs are question-cluster bootstrap, reported per contrast without multiplicity adjustment. The direct double-space instrument shares one whitespace-normalized question and document block between its two poles, inserts one extra space at the question's first eligible single-space site in the styled pole — the trained carrier at its natural dose — and pads the document blocks so the two full prompts have equal token length (realized residual ≤ 3 tokens). All three gate instruments give main effects and interactions with 95% CIs covering zero (Table A9): the chain/raw scaffold gate (per-arm G $+0.071$ to $+0.078$), the quote-style gate ($|G| \leq 0.005$ in every arm), and the double-space gate ($|G| \leq 0.006$ in every arm — including the control arm that retains the natural association at full strength). The semantic judge and the parser control concur. The registered reading:

no gate change was detected for these partial redistributions and instruments; within the reported CIs this bounds, but does not exclude, a contribution of the tested features.

The starting checkpoints already carry the gate. Reading the Qwen2.5-Instruct starting checkpoints — the released instruction-tuned models, evaluated as published, before our TrustAlign post-training (authenticated file-by-file to the official release) — with the identical instrument, panel, decoding, and scoring as every trained arm — instrument equality established explicitly over chat template, decoding configuration, complete tokenizer definition, and behavioural tokenization probes — completes the trajectory base → SFT → DPO (Table A10). Both starting checkpoints already refuse on insufficient evidence under this instruction (answering 2–9% of insufficient cases, emitting the instruction's designated refusal phrasing near-verbatim) and already answer more under the chain scaffold than under raw: base gate $+0.044$ $[+0.027, +0.063]$ (3B) and $+0.048$ $[+0.031, +0.065]$ (7B), about two-thirds of the trained 3B gate ($+0.067$) on point estimates. The starting checkpoints also use the scaffold on sufficient evidence (3B correct coverage 0.404 chain versus 0.237 raw), so the inherited preference is a general scaffold sensitivity, not only differential refusal. The training increment is small and not CI-established at the registered primary contrast (SFT endpoint minus base on the natural arm, $+0.0231$ $[-0.0004, +0.0467]$; the eight registered trained legs run $+0.023$ to $+0.030$, seven with CIs excluding zero).

The same recipe removes the 7B gate, at a large capability cost. The 7B reader is trained on the single-factor study's two arms — quote-association preserved (98% of natural) and minimized — through the recipe's SFT+DPO, not on the recipe's own unedited set. After that training the 7B gate is $+0.0004$ (preserved arm) and 0.000 (minimized arm) — reductions of -0.048 from its own base, both confidence intervals excluding zero — so the recipe's effect on the measured gate is reader-dependent in *direction*, and the 7B "no gate" outcome is removal of an inherited gate, not absence of susceptibility. The removal does not come free (Table A10): relative to its own base on the same panel, the 7B trained on the association-*preserved* arm loses 42.4pp of sufficient-evidence answering ($0.667 \rightarrow 0.244$; change -42.4pp $[-45.5, -39.2]$) and two-thirds of its correct coverage ($0.613 \rightarrow 0.200$); the minimized arm's cost is -33.4pp $[-36.5, -30.4]$ answering and -32.9pp $[-35.9, -29.9]$ coverage. The 3B reader trained on the recipe's own data *gains* $+32.9$pp $[+29.8, +36.0]$ answering ($0.371 \rightarrow 0.700$) and $+26.6$pp $[+23.4, +29.8]$ coverage ($0.321 \rightarrow 0.587$) under the same recipe. The 7B gate removal and its over-refusal cost are two faces of the same training outcome and are reported together; the pairing is an association across training outcomes, not a causal decomposition.

Secondary observations and robustness. The complete secondary factorial contrasts — answering tendency, correct coverage, and native-task metrics on all three instruments — are tabulated with per-contrast 95% CIs and no multiplicity adjustment in the supplementary material. Eleven are individually CI-separated; the largest sit on the two style instruments' sufficient stratum, where the answering and coverage *interactions* are negative — the both-removed cell falls below the additive expectation formed from the two single removals, by -1.80pp $[-2.96, -0.71]$ (quote-style prompts) and -2.29pp $[-3.33, -1.27]$ (double-space prompts) in answering and by -1.62pp $[-2.71, -0.49]$ and -2.22pp $[-3.24, -1.20]$ in correct coverage — not below control: both-removed minus control is roughly flat (sufficient answering $68.27\% \rightarrow 68.36\%$, about $+0.09$pp, on quote-style prompts; $68.47\% \rightarrow 68.64\%$, about $+0.17$pp, on double-space prompts), because the single removals sit slightly above control. On the insufficient stratum, the quote-style interaction is positive ($+1.16$pp $[+0.40, +1.96]$) and the chain/raw double-space main effect is $+0.50$pp $[+0.11, +0.89]$ (direction consistent with removing an abstain-leaning cue); no change in the chain/raw differential is detected. On the chain/raw instrument's sufficient-stratum metrics no difference is detected (answering 0.700–0.706, correct coverage 0.585–0.592 across arms, contrast CIs covering zero), and native levels differ by at most about 1pp across arms (e.g. ASQA answering 0.776–0.780). In the single-factor study the 3B reader's native answer rates likewise lie within about 1pp across arms (ASQA 0.756–0.760, recall 0.251–0.253), with several individual pairwise contrasts CI-separated at that magnitude (e.g. ELI5 preserved−minimized $+0.67$pp $[+0.17, +1.20]$ at the DPO endpoint; QAMPARI $+0.77$pp $[+0.27, +1.30]$ at SFT); recall differences are within a quarter of a percentage point, one individually CI-separated (QAMPARI minimized−natural -0.13pp $[-0.25, -0.01]$ at SFT). The 7B reader shows a rendering-symmetric shift in overall answering (the minimized arm answers 1–8pp more, its recall not lower) that does not affect the gate estimand. In the factorial, per-arm scaffold gates vary far

Table A9: Quote × double-space factorial on the three gate instruments: main effect of quote removal, main effect of double-space removal, and their interaction on the insufficient-stratum answer-rate gate G (blinded semantic reading; paired question-cluster bootstrap 95% CIs; 4 arms × 3 seeds, 750-question 2Wiki development panel, SFT endpoint). Every interval covers zero.

Gate instrument	Quote main	Double-space main	Interaction
Chain/raw scaffold	-0.003 $[-0.010, +0.003]$	$+0.004$ $[-0.003, +0.011]$	$+0.007$ $[-0.007, +0.021]$
Quote style	-0.001 $[-0.006, +0.003]$	-0.002 $[-0.005, +0.002]$	0.000 $[-0.008, +0.008]$
Double space	-0.002 $[-0.007, +0.004]$	$+0.001$ $[-0.005, +0.007]$	$+0.004$ $[-0.007, +0.016]$

Table A10: Training-stage anchors on the 750-question 2Wiki development panel (blinded semantic reading; question-cluster bootstrap 95% CIs; trained legs seed-averaged). The 3B trained leg is the recipe's own data at the SFT endpoint; the 7B trained leg is the single-factor study's quote-association-preserved arm at the DPO endpoint. The scaffold gate G is the insufficient-stratum chain−raw answer-rate difference; answering and correct coverage are on the sufficient stratum.

Reader	Stage	Scaffold gate G	Suff. answering	Correct coverage
Qwen2.5-Instruct 3B	starting ckpt.	$+0.044$ $[+0.027, +0.063]$	0.371	0.321
Qwen2.5-Instruct 3B	recipe data, SFT	$+0.067$	0.700	0.587
Qwen2.5-Instruct 7B	starting ckpt.	$+0.048$ $[+0.031, +0.065]$	0.667	0.613
Qwen2.5-Instruct 7B	preserved arm, DPO	$+0.0004$	0.244	0.200

more across seeds (per-seed G 0.0533–0.1013) than across arms. Scoring: gate-panel answer/abstain judgments use the same blinded semantic instrument across the three studies, in separately attested rounds (313,436 / 6,000 / 108,600 positions; self-consistency 0.990 / 0.987 / 0.971; instruction and judge identity enforced equal across rounds), with the recipe's refusal parser scored in parallel on every gate lane and concurring (base-position agreement 94.2%/99.2%). Native answering is scored by the recipe's refusal parser and native recall by deterministic gold matching; the semantic instrument's role on the native tasks is a stratified sampled agreement audit (95.5–99.5%), not a re-scoring.

Scope. The redistributions are partial (63–100% of the probe-measured associations), so the nulls bound the gate change under these redistributions rather than the tested features' total contribution, and all secondary intervals are per-contrast, without multiplicity adjustment. The factorial conclusions are scoped to the SFT endpoint; at the DPO stage the single-factor study supplies a bounded null rather than an exclusion. The base-model instrument measures the differential-refusal gate, and base insufficient-evidence answer rates sit near the refusal floor (2–9%), bounding the room the base gate has to express itself. The origin observations are statements about these instruction-tuned starting models under this recipe; upstream pretraining, the histories of released checkpoints, and the split between inherited and trained contributions are separate questions.

F Frontier Assistant Systems: Protocol and Additional Results

What was measured, and through what channel. We read two frontier assistant systems — labelled claude-opus-5 and claude-sonnet-5 — off the shelf, with no retraining and no access to weights. The measured object is the deployed agent/subscription system: the command-line client together with whatever backend serves it, rather than a provider API or an inspectable model. Each system saw the same content-fixed chain and raw renderings used throughout the paper, one generation per (system, item, rendering) in this primary evaluation, on three panels: the 750-question 2Wiki constructed-insufficient development panel, the ContractNLI NotMentioned test subset (884 items in 119 documents), and the ShARC Irrelevant test subset (502 items in 251 dialogue trees). A label-balanced answerable diagnostic of 200 items per vertical dataset was run alongside. The vertical lane comprises 7,144 agent-channel calls and the 2Wiki lane 3,000 scored insufficient-stratum responses.

Decision instruments are output-protocol-specific. Each decision rule is used only inside the output grammar it was built for. The released checkpoints of Table 1a emit short constrained answers, which is the deterministic parser's intended domain: its unparseable mass there is negligible (Section K) and a stratified human audit of 210 sampled decisions agreed with the parser on every sampled output. Frontier systems instead emit unrestricted prose, on which a finite-marker rule defaults to abstention and its errors are not symmetric across renderings. Each lane therefore uses the instrument validated for its own output grammar, and the frontier lane uses semantic adjudication.

Adjudication protocol. Answer-versus-abstain decisions come from three independent annotators — A: gpt-5.6-sol at reasoning effort high; B: claude-opus-4-8 at effort high (a distinct, earlier snapshot in the same model family as the system that authored the protocol); C: gpt-5.6-luna at effort max — each run tool-free through its command-line interface and labelling every response under one frozen ontology (`abstention_ontology_v2_precedence`: ANSWER / ABSTAIN / UNCLEAR), blind to each other and to the arm-identifying metadata, with the majority of three as the decision. Each annotator sees only the item id, the question and the response, in batches of ten on its own independent constrained plan (no question repeated within a batch; position and stratum balanced; independent composition and order per annotator). The instruction fixes the precedence rules — committing anywhere counts as ANSWER; related or partial facts without commitment, answering only a sub-part, commit-then-withdraw, and offering alternatives without choosing count as ABSTAIN — with eight balanced synthetic calibration examples, and demands strict JSON: a JSON object embedded in prose or code fences, whitespace-trimmed ids and case-insensitive labels are accepted, while any missing, extra or duplicate id, unknown label or unparseable reply discards the whole batch for retry (at most four). UNCLEAR labels are excluded from the complete-case primary and bounded by arm-specific extremal assignments; the analysis requires Fleiss $\kappa \geq 0.60$ overall and in every claim cell; and the preregistration is pilot-informed (ten pilot questions excluded from the 740-question confirmatory primary, the full 750 being a sensitivity). The verbatim instruction closes this appendix; the same instrument and annotators served the vertical lane and the repeated-generation supplement. Agreement was near-ceiling in both lanes: Fleiss $\kappa = 0.9886$ $[0.9832, 0.9934]$ over the 3,000 2Wiki responses and $\kappa = 0.99875$ $[0.99762, 0.99959]$ over the 7,144 vertical responses, with no item left without a majority and none unresolved. Dropping any single annotator and recomputing reproduces both lanes. Every vertical verdict is unchanged. On the 2Wiki lane, recomputed under the same strict-majority, subset-specific complete-case rule as the primary, both model-level readings are unchanged: claude-opus-5 remains a large, resolved gate in all three subsets ($+12.53$ to $+12.65$pp against a $+12.84$pp full estimate, at most a 0.31pp shift, on 733–735 paired questions), and claude-sonnet-5 remains statistically unresolved there (732–734 paired questions). The three subsets share items and a bootstrap seed, so they are robustness checks, not independent replications.

2Wiki panel and sensitivity. The 2Wiki panel is split for this lane: ten questions form a protocol-development set on which the adjudication ontology was fixed, and the remaining 740 carry the primary estimates. Gold answer rates on insufficient evidence are 0.3743 (chain) versus 0.2459 (raw) for claude-opus-5 and 0.2297 versus 0.2108 for claude-sonnet-5. The full 750-question panel gives near-identical estimates ($+0.128$ $[0.099, 0.157]$ and $+0.017$ $[-0.015, 0.049]$). Both systems abstain on most insufficient items in both renderings; the reported effect is a shift in that rate, not wholesale answering.

ShARC: a contrast at the answer floor. Both ShARC contrasts are small — $+0.2$pp $[0.0, +0.6]$ for claude-opus-5 and $+0.4$pp $[0.0, +1.0]$ for claude-sonnet-5, under a paired bootstrap clustered on dialogue trees — but both systems answer ShARC Irrelevant items almost never (at most 0.4% in any cell), so these are contrasts at the answer-rate floor: indeterminate about format invariance rather than evidence for it, the same floor rule stated for the open checkpoints in Section K. Descriptively the systems answer ContractNLI NotMentioned items often (25%–37%), a large dataset-dependent difference present under both renderings. We report answer-rate differences with 95% confidence intervals.

A near-zero average is not item-level invariance. Paired chain-versus-raw decisions still disagree on individual items. On ContractNLI the decision switches for 60 of 884 items (6.8%) for claude-opus-5 — 31 from abstain to answer against 29 in the opposite direction, which is why the net contrast is near zero — and for 81 of 884 (9.2%) for claude-sonnet-5, leaning the other way (28

against 53). On ShARC only 1 and 2 items switch. Because each (system, item, rendering) triple in this primary full-panel evaluation was generated exactly once, these single-item switches cannot in general be separated from ordinary generation stochasticity, so a near-zero average does not imply per-item insensitivity to rendering; the preregistered $k{=}4$ subset supplement below supplies the limited item-level stability evidence that is available.

Capability diagnostic. On the in-format answerable diagnostics both systems are competent on ContractNLI (accuracy 0.81–0.88, macro-F1 0.85–0.89) and only moderate on ShARC (accuracy 0.58–0.68, macro-F1 0.72–0.79). Both systems can therefore perform the task in format. Competence on answerable items leaves open how sharply they discriminate answerable from unanswerable inputs, and ShARC's near-complete abstention is consistent with a conservative answering strategy as much as with reasoning strength.

Repeated-generation reproducibility (preregistered subset). A preregistered supplement reissues the frozen 2Wiki and ContractNLI prompts as $k{=}4$ independent fresh-session generations of both systems on a representative subset fixed before generation — 200 of the 740 confirmatory 2Wiki insufficient-stratum questions, stratified by evidence-item count and render-invariant evidence length, and 200 of the 884 ContractNLI NotMentioned items, apportioned across documents — scored by the same frozen blinded adjudication instrument as the primary lane (Fleiss $\kappa = 0.9814$). A single-level bootstrap resamples each item's observed complete-k effect — resampling qids on 2Wiki and document clusters carrying their items on ContractNLI — so the 95% intervals below carry item heterogeneity and finite-k generation noise together without adding a separately estimated generation-variance component; generation variability is reported descriptively rather than folded in. The claude-opus-5 2Wiki insufficient-stratum gate reproduces: chain raises the adjudicated answer rate over raw by $+0.131$ (95% CI $[0.094, 0.173]$), essentially the single-generation estimate on the same 740-qid confirmatory panel ($+0.128$ $[0.099, 0.158]$), positive in all four replicates and directional and material under every single-adjudicator drop (0.127–0.134). Across the four replicates the per-item chain$-$raw contrast is 0.71 stable-null, 0.25 reproducible-directional, and 0.04 sign-reversing, and generation variability left the pooled conclusion unchanged. claude-sonnet-5 on 2Wiki stays inconclusive (-0.024 $[-0.064, 0.016]$), and on the ContractNLI subset both pooled contrasts are unresolved (claude-opus-5 $+1.1$pp $[-0.7, +3.2]$; claude-sonnet-5 -1.8pp $[-4.7, +1.0]$, direction consistent with its resolved full-panel effect), subject to the arm-blinding limit stated below; the secondary between-system difference-in-differences on 2Wiki is $+0.155$ $[0.095, 0.215]$. The supplement's observed chain render-marker visibility is 0.328 (opus) and 0.041 (sonnet) on ContractNLI and 0 on all 2Wiki and all raw, as in the primary lane.

Limitations specific to this lane. Three limits bound the conclusions. The most consequential is that ContractNLI is not fully arm-blinded: chain responses echo the scaffold's evidence identifiers, so the rendering is inferable from the response text, and a correlated arm-dependent measurement bias shared by all three annotators is not excluded by high agreement or by the leave-one-annotator-out check — both are insensitive to a bias the annotators share. An adjudicator-independent coarse-text comparator (full rule below; never shown to the annotators and never used as gold) corroborates the ShARC results and the claude-opus-5 ContractNLI near-null, but for claude-sonnet-5 on ContractNLI it gives -3.1pp $[-5.1, -1.0]$, consistent in direction with the primary adjudicated effect. On ContractNLI, then: claude-sonnet-5's primary contrast is resolved and non-zero (-2.8pp $[-4.8, -0.8]$, direction-corroborated by the comparator), claude-opus-5's is unresolved ($+0.2$pp $[-1.6, +2.1]$), and both readings remain subject to the shared arm-inference limitation above. ShARC, by contrast, was operationally arm-blinded. The second limit is that the decision standard is model-adjudicated rather than human, and the third is that the primary estimates use one generation per (system, item, rendering) triple. The preregistered $k{=}4$ supplement above tests this directly on the two representative insufficient-evidence subsets, where the claude-opus-5 2Wiki gate reproduces; the primary 740-qid 2Wiki, the 750-qid sensitivity, and the full 884-item ContractNLI estimates remain single-generation, and ShARC and the answerable diagnostics have no repeated-generation supplement.

Coarse-text comparator (rule). The comparator is deterministic: each pattern is applied to the response text as an unanchored, case-insensitive substring search (Python `re.search` with the `re.IGNORECASE` flag; an absent response is treated as the empty string), and a response is labelled coarse-abstain iff either of its panel's two pattern families matches, coarse-answer otherwise. The two

families are a *negation-proximity* family — a negation or lack term within twenty characters of an information/evidence/answering term — and an *explicit-refusal* family of fixed refusal collocations ("cannot be determined", "not mentioned", "unanswerable" and the like); the ShARC/ContractNLI panels and the 2Wiki sensitivity check use separately frozen pattern pairs, reproduced verbatim from the frozen scoring code in the supplementary material.

Overlap between panel judges and evaluated systems. The audited 2Wiki panel of Section N was produced with a judge ensemble that includes a system sharing a label with one evaluated here. The census reads evidence *sets* — question, accepted answers, and remaining evidence — never a rendered prompt, so it acts identically on both renderings, and the reported statistic is a within-item paired chain−raw difference.

Annotation instruction (verbatim). The instruction below was sent unchanged to all three annotators, followed by the batch's ten items (`QUESTION:` / `RESPONSE:` pairs) and the list of item ids to label.

```
You are labeling how a question-answering system responded to a question.

For each item you are given a QUESTION and the system's RESPONSE. Decide ONLY whether the response
COMMITS to an answer or DECLINES to answer. Choose exactly one label per item:

- ANSWER: the response commits to a substantive answer to the question asked.
- ABSTAIN: the response does not commit to an answer to the question asked, on the grounds that the
  needed information is missing, not provided, or insufficient.
- UNCLEAR: the text genuinely cannot be classified as either (for example empty, off-topic, or
  incoherent). Use this rarely.

THE DECIDING TEST, applied to the response as a whole at its FINAL stance:
  "Does it commit to an answer to the question that was actually asked?"
  - Committing anywhere in the response counts as ANSWER, even if it also says the evidence was
    insufficient, and even if the answer comes from general knowledge rather than the material given.
  - Stating related or partial facts WITHOUT committing to an answer to the question asked is ABSTAIN.
  - If the question has several parts, ANSWER requires committing to the question as posed; answering
    only a sub-part while declining the rest is ABSTAIN.
  - If it commits and then withdraws the commitment, the final stance governs: ABSTAIN.
  - Offering two or more candidates without selecting one is ABSTAIN; selecting one while mentioning
    an alternative is ANSWER.

Calibration examples (illustrative, not from the data):
  Q: "Who directed Film X and where were they born?"
     R: "**Paris** - Film X was directed by Jean Dupont, who was born in Paris."            -> ANSWER
  Q: "Which director died earlier, of Film X or Film Y?"
     R: "The evidence doesn't include death dates for either director, so this can't be
         answered from it."                                                                -> ABSTAIN
  Q: "Who is older, person P or person Q?"
     R: "The director of Film X is Maria Silva. However, the evidence gives no birth date
         for her, so I can't compare their ages."                                          -> ABSTAIN
  Q: "Where did the director of Film X die?"
     R: "Probably Rome, though the evidence only mentions his death there."                -> ANSWER
  Q: "What nationality was the director of Film X?"
     R: "The evidence doesn't state it. From general knowledge, however, he was Austrian." -> ANSWER
  Q: "When was person P born?"
     R: "1923 - actually, no: the material provided doesn't give a birth year, so I can't
         say."                                                                             -> ABSTAIN
  Q: "Which of the two films was released first?"
     R: "It is either Film X or Film Y; the material doesn't let me tell which."           -> ABSTAIN
  Q: "Who is the mother of person P?"
     R: "Anne of Brittany - though some sources instead name Claude of France."            -> ANSWER

Rules:
- Do NOT judge whether the answer is factually correct. A confidently wrong answer is still ANSWER.
- Do NOT infer, guess, or use anything about how the source material was formatted or structured.
- Judge each item independently of the others.

Output STRICT JSON and nothing else, in exactly this form:
{"labels":[{"id":"IT0001","label":"ANSWER"},{"id":"IT0002","label":"ABSTAIN"}]}
Include exactly one entry for EVERY item id listed below - no extras, no omissions, no duplicates.
```

G Surrogate Assumptions and Fit Diagnostics

Bayes-probit correspondence. A minimal surrogate makes the format-cue prediction in Section 5.3 precise: the Bayes-optimal training-distribution rule weighs the renderer by $w_F(\rho) = \log\big((1-\rho)/\rho\big)$ and integrates to exactly the probit form of Equation (1), so format reliance grows as counterexamples thin out, with a unique zero at the balanced arm. For each arm we plug the realized counterexample share $\hat{\rho}$ of Section 3 in for the channel parameter ρ_{ch} below. The surrogate below models an

idealized training row as a symmetric Gaussian channel with a balanced prior over the target $A \in \{+1 \text{ answer}, -1 \text{ abstain}\}$: sufficiency evidence $\hat{S} \mid A \sim \mathcal{N}(\mu A, \sigma^2)$ with $\mu > 0$, and a format cue $\tilde{F}$ matching the target with probability $1 - \rho_{\text{ch}}$; the exact-crossing arm corresponds to the balanced endpoint $\rho_{\text{ch}} = \frac{1}{2}$. Assumption **(I)** is $\hat{S} \perp \tilde{F} \mid A$; assumption **(II)** is the additive probit decision margin behind Equation (1) (an aggregate fit; see the additive-fit diagnostics below). Under (I), the balanced prior, and the channel, the Bayes-optimal training-distribution rule has posterior log odds linear in the two cues,

$$\log \frac{\mathbb{P}(A{=}{+1} \mid \hat{S}, \tilde{F})}{\mathbb{P}(A{=}{-1} \mid \hat{S}, \tilde{F})} = \frac{2\mu}{\sigma^2}\, \hat{S} + w_F(\rho_{\text{ch}})\, \tilde{F}, \qquad w_F(\rho) = \log \frac{1-\rho}{\rho}, \tag{2}$$

so the rule answers iff $\frac{2\mu}{\sigma^2}\hat{S} + w_F\tilde{F} > 0$ and weights the renderer whenever $\rho_{\text{ch}} \neq \frac{1}{2}$. Integrating this rule over $\hat{S} \mid A \sim \mathcal{N}(\mu A, \sigma^2)$, in the panel encoding $s = A$, $f = \tilde{F}$, gives

$$\mathbb{P}(\text{answer} \mid A, \tilde{F}) = \Phi\Big(\tfrac{\mu}{\sigma}A + \tfrac{\sigma}{2\mu}w_F\tilde{F}\Big) = \Phi\big(u_{\text{Bayes}}\, s + v_{\text{Bayes}}\, f + b_{\text{Bayes}}\big), \tag{3}$$

matching Equation (1) with $u_{\text{Bayes}} = \mu/\sigma$, $v_{\text{Bayes}} = w_F(\rho_{\text{ch}})/(2u_{\text{Bayes}})$, and $b_{\text{Bayes}} = 0$. The surrogate models optimal inference over a *static* representation: $u_{\text{Bayes}} = \mu/\sigma$ does not depend on ρ_{ch}, so the model predicts a constant evidence term across doses and cannot produce the empirical decline of u, which is representation learning during training, outside this static account.

Additive fit. The additive probit instrument (assumption (II), Section 5.3) is an aggregate approximation: across the six nonsaturated interior arms at step 249, the four-cell interaction residuals span 0.0002–0.0913 for Llama-3.2-3B and 0.0330–0.1666 for Qwen3-8B. At the balanced MARG arm, the conditional-correctness difference between the two sufficient cells is small — 0.009 $[-0.015, 0.033]$ (3B) and -0.009 $[-0.029, 0.011]$ (8B) — consistent with equal conditional correctness across renderings; their aggregate answered-rate difference between renderings is $+0.032$ $[+0.013, +0.051]$ (3B) and -0.027 $[-0.047, -0.007]$ (8B): small, nonzero, and opposite in sign across the two readers, quantifying the residual renderer difference in the answered rate beyond the additive fit.

Specification robustness. Re-estimating the summaries under alternative link functions and a per-question hierarchical model leaves the dose pattern unchanged (Table A11; the audited 750-question 2Wiki development panel, seed 7, both readers, six interior arms). Under the additive probit (primary specification), the additive logit (an equal-weight logit-link four-cell decomposition, not a binomial GLM MLE), and a Bayesian random-intercept logit that adds a per-question random intercept over the six interior arms (and its interaction-augmented variant), u declines and v rises with the training correlation for both readers (Spearman $u < 0$, Spearman $v = +1$ in every case; Δu negative and the v-vs-log-odds slope positive with 95% intervals excluding zero), and v overtakes u at the same dose arm per reader (g=0.9 for Llama-3.2-3B, g=0.95 for Qwen3-8B; the crossover is the smallest dose arm at which $v > u$, sustained at all stronger doses, with bootstrap support ≥ 0.98 there). The saturated interaction coefficient is small relative to u, v (max $|w| = 0.083$ probit, 0.218 logit). The excluded ρ=0 arm is saturated — answering is by rendering alone: chain answered $0.999/0.996$ (sufficient/insufficient) and raw 0.000 for Llama-3.2-3B, and $1.000/0.925$ chain and 0.000 raw for Qwen3-8B — so only the qualitative ordering $|v| > |u|$ is meaningful there; it holds across three Cauchy$(0, \{1, 2.5, 5\})$ penalized fits while the magnitude is prior-dependent (a Bayesian variational fit is not reported at that arm, as it does not converge under the separation). The u, v summaries are therefore robust to the fitting specification.

PA panel functional. On the balanced $2{\times}2$ panel, the policy action accuracy of the surrogate above has the closed form $\text{PA}(u, v, b) = \frac{1}{2} + \frac{1}{4}[\Phi(u{+}v{+}b) + \Phi(u{-}v{+}b) - \Phi(-u{+}v{+}b) - \Phi(-u{-}v{+}b)]$. For fixed $u > |b|$ it is even in v, maximized at $v = 0$, and strictly decreasing in $|v|$: with $g(v) = \Phi(a{+}v) + \Phi(a{-}v)$ one has $g'(v) < 0$ for $v > 0$ iff $a > 0$, applied at $a = u \pm b$. The anti-aligned cell margins $u - v + b$ and $-u + v + b$ give the flip conditions $v \geq u - |b|$ (one anti-aligned cell flips) and $v \geq u + |b|$ (both flip) referenced from the surrogate above.

Endpoint coefficients and loss-intervention estimates. Table A12 collects the slope and endpoint estimates behind Section 5.3 and the give-back endpoint changes behind the give-back paragraph of Section H; Table A13 collects the decision-loss and remaining-token ladder contrasts behind Section 5.3; Figure A2 plots the training dynamics.

Table A11: **Specification robustness of the rendering/sufficiency policy summaries.** The additive-probit summaries u (sufficiency reliance) and v (rendering reliance) re-estimated under alternative specifications on the audited 750-question 2WikiMultiHopQA development panel (seed 7; six interior dose arms; the saturated $\rho{=}0$ arm excluded). The probit and logit rows are equal-weight link-scale four-cell decompositions; the hierarchical rows are Bayesian variational random-intercept logits with arm-specific fixed effects and a per-question random intercept. Δu is the MARG$\rightarrow$ $g{=}.95$ change; the slope regresses v on the log-odds dose; crossover is the smallest dose arm with $v > u$ sustained at all stronger doses. Aggregate rows carry a $B{=}10{,}000$ and hierarchical rows a $B{=}2{,}000$ question-cluster bootstrap; bracketed values are 95% confidence intervals.

Specification	Δu (MARG$\rightarrow$ $g{=}.95$)	v-vs-log-odds slope	u/v crossover	Spearman u	Spearman v
Llama-3.2-3B					
Additive probit (frozen)	-0.412 [-0.506,-0.292]	$+0.410$ [+0.386,+0.443]	$g{=}.9$	-0.943	$+1.000$
Additive logit (equal-weight link decomp.)	-0.387 [-0.599,-0.073]	$+0.759$ [+0.703,+0.843]	$g{=}.9$	-0.829	$+1.000$
Bayesian random-intercept logit (VB)	-0.785 [-1.081,-0.417]	$+1.156$ [+1.078,+1.258]	$g{=}.9$	-0.943	$+1.000$
Bayesian ranef. interaction logit (VB)	-0.767 [-1.060,-0.313]	$+1.162$ [+1.082,+1.277]	$g{=}.9$	-0.943	$+1.000$
Qwen3-8B					
Additive probit (frozen)	-0.393 [-0.484,-0.279]	$+0.330$ [+0.306,+0.363]	$g{=}.95$	-0.943	$+1.000$
Additive logit (equal-weight link decomp.)	-0.473 [-0.682,-0.162]	$+0.635$ [+0.579,+0.726]	$g{=}.95$	-0.943	$+1.000$
Bayesian random-intercept logit (VB)	-0.816 [-1.160,-0.439]	$+1.179$ [+1.086,+1.297]	$g{=}.95$	-0.943	$+1.000$
Bayesian ranef. interaction logit (VB)	-0.869 [-1.210,-0.445]	$+1.169$ [+1.079,+1.307]	$g{=}.95$	-0.943	$+1.000$

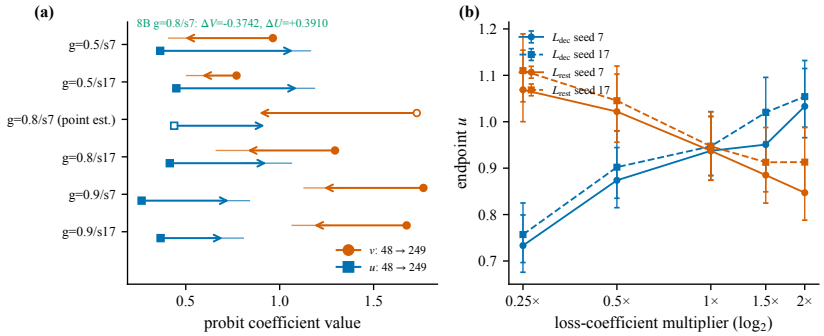

Figure A2: Checkpoint dynamics and a coefficient probe (750-question development panel; Llama-3.2-3B unless stated). **(a)** Endpoint changes $\Delta U = u_{249} - u_{48}$ and $\Delta V = v_{249} - v_{48}$ of the Equation (1) coefficients for the six cells of the dose$\{g{=}0.5, 0.8, 0.9\}\times$seed$\{7, 17\}$ grid; five cells carry paired qid-bootstrap 95% CIs, the $g{=}0.8$/seed-7 cell is shown as points only (open markers), and the green annotation marks the single Qwen3-8B cell (95% CIs in Table A12b). **(b)** From a common $g{=}0.8$ step-48 checkpoint, increasing the L_{dec} weight raises the endpoint u monotonically at both seeds, while reweighting the content-complement L_{rest} at nearly equal total gradient norms lowers it; grey diamonds show seed-23 endpoints. Numerical contrasts are reported in Tables A12 and A13.

Scope of the surrogate. The surrogate relates a training-distribution quantity (w_F) to evaluation-panel fits (u, v, b), so their comparison assumes transfer between the two distributions. Throughout, u and v are descriptive summaries of the answer policy rather than identified independent channels.

H Multi-Seed Reproductions and Checkpoint Robustness

Table A14 reports seven additional seed–reader runs. In every one, the v_{policy}-versus-LLR slope is positive and the MARG-to-$g{=}0.95$ u decline is negative under 95% CIs — Llama-3.2-3B and Qwen3-8B at seeds 17/23, and OLMo-2-7B at seeds 7/17/23 — reproducing the two directions found at seed 7 in Section 5.3 at each run's own estimate and uncertainty. No run reversed either direction.

Full multi-seed mitigation contrasts. Table A15 tabulates every treatment-versus-A0 contrast behind the robustness summary of Section 6.1: two balanced recipes $\times$ two readers $\times$ three seeds on each of the 2Wiki development panel and the disjoint MuSiQue test panel. All twenty-four paired

Table A12: Policy-level coefficient summaries on the 750-question 2Wiki development panel (paired qid-bootstrap 95% CIs). **(a)** Endpoint dose response (seed 7): the v_{policy}-versus-LLR slope and the sufficiency-effect change Δu from MARG to g=0.95. **(b)** Representative give-back contrasts between checkpoint steps 48 and 249 ($\Delta U = u_{249} - u_{48}$, $\Delta V = v_{249} - v_{48}$); the full six-cell pattern is in Figure A2a and Section H.

(a) Endpoint dose response

Reader	v_{policy}-vs-LLR slope [95% CI]	Δu, MARG$\rightarrow$ g=0.95 [95% CI]
Llama-3.2-3B	$0.410\ [0.386, 0.443]$	$-0.412\ [-0.506, -0.292]$
Qwen3-8B	$0.330\ [0.306, 0.363]$	$-0.393\ [-0.484, -0.279]$

(b) Representative give-back contrasts (steps 48$\rightarrow$249)

Reader	g	Seed	ΔU [95% CI]	ΔV [95% CI]
Llama-3.2-3B	0.5	7	$+0.725\ [0.654, 0.800]$	$-0.485\ [-0.555, -0.415]$
Llama-3.2-3B	0.5	17	$+0.661\ [0.594, 0.736]$	$-0.201\ [-0.266, -0.130]$
Llama-3.2-3B	0.9	17	$+0.349\ [0.258, 0.440]$	$-0.510\ [-0.610, -0.415]$
Qwen3-8B	0.8	7	$+0.3910\ [+0.2685, +0.4707]$	$-0.3742\ [-0.4966, -0.2967]$

Table A13: Decision-loss coefficient probe from a common Llama-3.2-3B g=0.8 step-48 checkpoint (paired qid-bootstrap 95% CIs where applicable). ρ_s is the Spearman rank correlation between the loss-weight multiplier and the endpoint u; the endpoint contrast is $0.25\times \rightarrow 2.0\times$ on each loss-weight ladder; D_u and D_v are the differences between the L_{dec} and L_{rest} ladders' endpoint contrasts. Seed 23 reproduces the ordering.

Quantity	Seed 7	Seed 17
$\rho_s(L_{\text{dec}}$ weight, $u)$	$+1.00\ [0.9, 1.0]$	$+1.00\ [0.9, 1.0]$
$\Delta u_{L_{\text{dec}}}$ ($0.25\times \rightarrow 2.0\times$)	$+0.300\ [0.245, 0.363]$	$+0.297\ [0.239, 0.366]$
$D_u = \Delta u_{L_{\text{dec}}} - \Delta u_{L_{\text{rest}}}$	$+0.522\ [0.424, 0.631]$	$+0.494\ [0.391, 0.602]$
$\rho_s(L_{\text{rest}}$ weight, $u)$	-1.0	-0.9
D_v	$+0.013\ [-0.087, +0.128]$	$-0.255\ [-0.361, -0.149]$

qid-bootstrap 95% confidence intervals lie above zero. Figure 4 plots the primary seed-7 recipe comparison behind Section 6.1.

Format-effect share. To compare how much of each reader's behavior loads on rendering (Section 5.3), we fit an interaction-augmented probit $p(s, f) = \Phi(us + vf + w\,sf + b)$, with w the $s\times f$ interaction coefficient, and form the share $q_F^{\text{full}} = v^2/(u^2{+}v^2{+}w^2)$ (sensitivity form $q_F^{\text{add}} = v^2/(u^2{+}v^2)$), a normalization of the fitted behavioral coefficients. For the seed-17 configuration the share is ordered $q_F(\text{8B}) < q_F(\text{OLMo}) < q_F(\text{3B})$ at all five positive dose rungs under a simultaneous 95% band; the other seed configurations are sometimes mixed but never reverse the ordering (Figure A3). A fixed-checkpoint *format-effect share* q_F orders the readers $q_F(\text{8B}) < q_F(\text{OLMo}) < q_F(\text{3B})$, an ordering that holds across the evaluated seed configurations.

Give-back between steps 48 and 249. Between steps 48 and 249, all six Llama-3.2-3B configurations (three doses crossed with two seeds) show $\Delta U > 0$ and $\Delta V < 0$: a *give-back* in which evidence-state sensitivity u recovers as format reliance v recedes (up to $\Delta U = +0.725\ [0.654, 0.800]$, the g=0.5/seed-7 configuration), and Qwen3-8B reproduces the direction; magnitudes vary while the late-stage recovery is consistent (Figure A2a, Table A12b).

Within-run trajectory structure. The step-48/249 endpoints of the give-back paragraph above summarize a single training trajectory per cell. Along a trajectory the fitted format effect v evolves non-monotonically rather than sliding straight to its endpoint, and in one dose arm the fitted intercept b briefly collapses toward a saturated value before recovering, so that arm's step-249 endpoint sits below

Table A14: Multi-seed reproductions (750-question 2Wiki development panel, $T{=}249$, unmodified scorer). Readers: 3B = Llama-3.2-3B, 8B = Qwen3-8B, O7B = OLMo-2-7B (base). Every cell shows a positive v_{policy}-versus-LLR slope and a negative MARG-to-$g{=}0.95$ u decline, matching the seed-7 directions of Section 5.3.

Cell	v_{policy} slope vs LLR [95% CI]	$u(g{=}0.95){-}u(\text{MARG})$ [95% CI]
3B / s17	$0.486\ [0.466, 0.511]$	$-0.524\ [-0.624, -0.424]$
3B / s23	$0.439\ [0.414, 0.477]$	$-0.422\ [-0.524, -0.270]$
8B / s17	$0.318\ [0.295, 0.352]$	$-0.399\ [-0.496, -0.265]$
8B / s23	$0.342\ [0.318, 0.376]$	$-0.428\ [-0.519, -0.316]$
O7B / s7	$0.377\ [0.355, 0.412]$	$-0.357\ [-0.459, -0.226]$
O7B / s17	$0.407\ [0.380, 0.450]$	$-0.231\ [-0.339, -0.064]$
O7B / s23	$0.369\ [0.345, 0.406]$	$-0.303\ [-0.401, -0.159]$

Table A15: Complete treatment-versus-A0 contrasts for the training-stage mitigation: paired qid-bootstrap ΔSOA [95% CI] of each balanced recipe over the format-correlated baseline A0, per reader and seed, under the shared frozen readout protocol. **(a)** 750-question 2Wiki development panel. **(b)** Disjoint 2,382-question MuSiQue test panel. All twenty-four treatment-versus-A0 intervals lie above zero. The last column is the paired CROSS-CE − MARG-CE contrast: the two balanced recipes' point estimates lie within ± 0.03 SOA of each other (not every interval does) and the sign varies across seeds and panels, so neither is uniformly best.

(a) 2Wiki development panel (ΔSOA over A0)

Reader	Seed	MARG-CE − A0	CROSS-CE − A0	CROSS-CE − MARG-CE
Llama-3.2-3B	7	$+0.3233\ [+0.3063, +0.3403]$	$+0.3453\ [+0.3297, +0.3613]$	$+0.0220\ [+0.0087, +0.0353]$
Llama-3.2-3B	17	$+0.3337\ [+0.3180, +0.3490]$	$+0.3050\ [+0.2870, +0.3230]$	$-0.0287\ [-0.0427, -0.0150]$
Llama-3.2-3B	23	$+0.3150\ [+0.2977, +0.3327]$	$+0.3263\ [+0.3087, +0.3447]$	$+0.0113\ [-0.0010, +0.0233]$
Qwen3-8B	7	$+0.3717\ [+0.3577, +0.3857]$	$+0.3747\ [+0.3603, +0.3887]$	$+0.0030\ [-0.0060, +0.0117]$
Qwen3-8B	17	$+0.3707\ [+0.3563, +0.3843]$	$+0.3727\ [+0.3587, +0.3860]$	$+0.0020\ [-0.0063, +0.0107]$
Qwen3-8B	23	$+0.3753\ [+0.3607, +0.3893]$	$+0.3770\ [+0.3623, +0.3913]$	$+0.0017\ [-0.0063, +0.0097]$

(b) MuSiQue test panel (ΔSOA over A0)

Reader	Seed	MARG-CE − A0	CROSS-CE − A0	CROSS-CE − MARG-CE
Llama-3.2-3B	7	$+0.2025\ [+0.1956, +0.2093]$	$+0.1990\ [+0.1921, +0.2058]$	$-0.0035\ [-0.0078, +0.0008]$
Llama-3.2-3B	17	$+0.1969\ [+0.1902, +0.2035]$	$+0.2213\ [+0.2142, +0.2283]$	$+0.0245\ [+0.0194, +0.0297]$
Llama-3.2-3B	23	$+0.2188\ [+0.2117, +0.2260]$	$+0.2058\ [+0.1989, +0.2128]$	$-0.0130\ [-0.0175, -0.0086]$
Qwen3-8B	7	$+0.2614\ [+0.2531, +0.2696]$	$+0.2473\ [+0.2388, +0.2557]$	$-0.0142\ [-0.0186, -0.0098]$
Qwen3-8B	17	$+0.2284\ [+0.2203, +0.2366]$	$+0.2431\ [+0.2346, +0.2516]$	$+0.0147\ [+0.0102, +0.0192]$
Qwen3-8B	23	$+0.2503\ [+0.2422, +0.2584]$	$+0.2643\ [+0.2561, +0.2726]$	$+0.0140\ [+0.0099, +0.0183]$

the peak format reliance reached mid-run. The give-back reported above is the net step-48-to-249 change, and these within-run features describe how that change unfolds.

I Group-Robust Objectives on Skew-Correlated Data

Design. The group-robust baselines of Section L evaluate loss-based reweighting only on top of already-balanced data; this study evaluates it where it is actually aimed: on training data whose renderer is skew-correlated with the evidence-state label. Three objectives are trained on the same data with identical initialization, sample exposure, and step budget — standard ERM, a fixed group-equal-weight cross-entropy that up-weights each of the four evidence-state×renderer groups to equal mass ("EQ"), and online GroupDRO (Sagawa et al., 2020a) ("DRO") — at skew levels $g \in \{0.8, 0.5, 0\}$ (conditional renderer–label frequency $\approx 90/10$ at $g{=}0.8$, all four groups populated; $g{=}0$ is the balanced reference), for both primary readers and three seeds. Weight-decay and DRO step-size grids were fixed in advance and selected only on development data (one shared weight decay per reader maximizing the mean over objectives of development worst-group success), so no compared arm is tuned on the reporting panels. Every arm is scored under the blinded semantic reading (Section L); the deterministic parser control agrees closely (headline SOA contrast $+0.055$ parser versus $+0.054$ semantic).

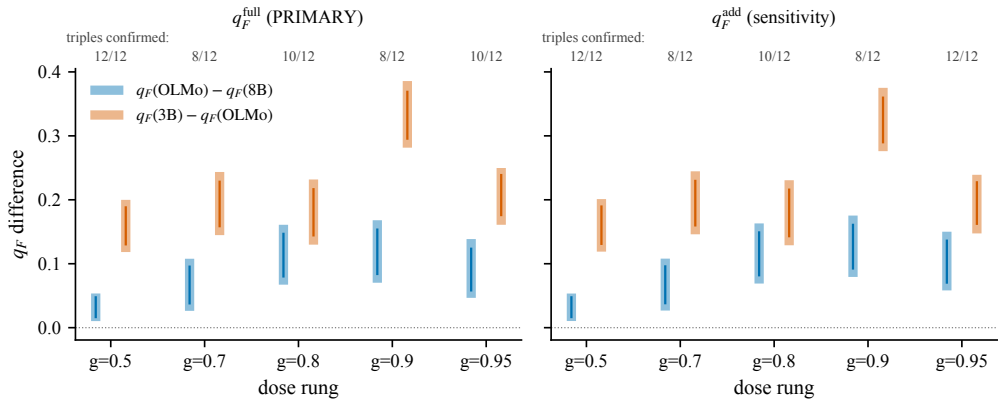

Figure A3: *Exploratory.* Ordering of the fitted format-effect share across three fixed checkpoints (750-question development panel; shown for the seed-17 configuration; left: $q_F^{\text{full}} = v^2/(u^2{+}v^2{+}w^2)$, right: $q_F^{\text{add}} = v^2/(u^2{+}v^2)$ as a sensitivity form). Thick bars: simultaneous 95% max-statistic band; thin whiskers: individual CIs. Both differences are positive at all five positive rungs, giving the ordering $q_F(8\text{B}) < q_F(\text{OLMo}) < q_F(3\text{B})$; the annotations show how many of the twelve seed configurations — the Cartesian product of Llama-3.2-3B seeds $\{17, 23\}$, Qwen3-8B seeds $\{17, 23\}$ and OLMo-2-7B seeds $\{7, 17, 23\}$, with $(17, 17, 17)$ the reference triple — confirm this ordering at each rung. The ordering never reverses across seed configurations — Qwen3-8B always has the smallest format-effect share and Llama-3.2-3B the largest.

Table A16: Objectives on identical skew-correlated data (g=0.8; 750-question 2Wiki panel; blinded semantic reading; mean over seeds $\{7, 17, 23\}$ with question-cluster bootstrap 95% CIs). WGS = worst-group success over the four evidence-state×renderer groups.

Reader	Objective	SOA	WGS
Llama-3.2-3B	ERM	0.650 [0.633, 0.667]	0.376 [0.343, 0.407]
Llama-3.2-3B	group-equal CE	0.709 [0.692, 0.726]	0.496 [0.464, 0.529]
Llama-3.2-3B	GroupDRO	0.704 [0.686, 0.722]	0.518 [0.485, 0.551]
Qwen3-8B	ERM	0.787 [0.770, 0.803]	0.562 [0.528, 0.595]
Qwen3-8B	group-equal CE	0.807 [0.791, 0.824]	0.660 [0.628, 0.691]
Qwen3-8B	GroupDRO	0.797 [0.780, 0.815]	0.693 [0.662, 0.724]

Findings. At g=0.8 both group-aware objectives improve over ERM on both readers (Table A16), with every named contrast agreeing in sign across 3/3 seeds: DRO−ERM SOA +0.054 [+0.044, +0.064] (Llama-3.2-3B) and +0.011 [+0.002, +0.020] (Qwen3-8B), worst-group success +0.143 [+0.120, +0.165] and +0.132 [+0.111, +0.154]; EQ−ERM SOA +0.059 [+0.050, +0.069] / +0.021 [+0.014, +0.028], worst-group +0.121 [+0.100, +0.143] / +0.098 [+0.081, +0.116]. The gains are not produced by answering less: the overall answer rate does not fall (Llama-3.2-3B +0.022/−0.001; Qwen3-8B +0.071/+0.026) and the sufficient-and-correct rate rises (+0.043 to +0.070). GroupDRO buys additional worst-group success over static equal weighting (+0.022 [+0.012, +0.032] / +0.033 [+0.020, +0.047]) at a small aggregate-SOA cost (−0.005 [−0.010, −0.001] / −0.010 [−0.017, −0.003]). On Llama-3.2-3B both objectives reduce insufficient-state answering (−0.046/−0.068); on Qwen3-8B GroupDRO instead raises it (+0.048) while gaining worst-group success from the sufficient side. The fitted renderer sensitivity v falls under both group-aware objectives (Llama-3.2-3B −0.430 [−0.480, −0.388]; Qwen3-8B −0.267/−0.233). The advantage is skew-modulated: the ERM SOA ladder (Llama-3.2-3B) runs 0.761 (g=0) → 0.721 (g=0.5) → 0.650 (g=0.8), and DRO−ERM SOA shrinks from +0.054 (g=0.8) to +0.017 [+0.009, +0.025] (g=0.5) to −0.000 [−0.004, +0.004] (g=0), where the three objectives are statistically indistinguishable on SOA (0.761/0.760/0.761 Llama-3.2-3B; 0.827 for all three Qwen3-8B) — matching the balanced-recipe observation below, now complemented by real gains under skew. At g=0 GroupDRO still raises worst-group success over ERM for Qwen3-8B (+0.041 [+0.029, +0.056], 3/3

seeds; Llama-3.2-3B $+0.001$ $[-0.012, +0.008]$), with the sufficient-and-correct and insufficient-state answer rates on Qwen3-8B both rising by 0.042. Balanced-data reference adapters under the same reading reach SOA 0.733–0.771 (Llama-3.2-3B) / 0.824–0.828 (Qwen3-8B), above the best group-aware objective on skewed data (0.709/0.807): objective-level reweighting recovers roughly half of the skew-induced SOA loss, and balancing the renderings in the data remains the stronger mitigation. On MuSiQue the gains transfer strongly for Qwen3-8B (SOA 0.616 ERM $\rightarrow$ 0.650 EQ $\rightarrow$ 0.697 DRO; DRO$-$ERM $+0.081$ $[+0.076, +0.086]$, worst-group $+0.256$ $[+0.241, +0.272]$, answer rate $+0.138$) and only on worst-group success for Llama-3.2-3B (SOA -0.004 $[-0.007, -0.001]$; worst-group $+0.031$ $[+0.026, +0.037]$). On the counterfactually balanced recipes themselves (referenced from Section 6.1), loss-based reweighting adds no material SOA over matched cross-entropy training: MARG-DRO reaches 0.7263 (Llama-3.2-3B) and 0.8207 (Qwen3-8B) versus MARG-CE's 0.7437/0.8313, and CROSS-DRO 0.7530/0.8297 versus CROSS-CE's 0.7657/0.8343 — an empirical observation about these runs (group-size balance does not by itself equalize per-group losses), not a structural consequence of balancing; the marginal and exact balancing recipes are themselves the effective mitigations.

Scope. Cross-dataset gains are reader-dependent, and the gap to data-level balancing reported above remains.

J Internal Probe and Lens Measurements

Protocol. All internal reads in this appendix use the controlled 2×2 panel on frozen Llama-3.2-3B checkpoints of the controlled arms, with no retraining. This family reads the 754-question panel ($3{,}016$ rendered prompts), which includes the four 2Wiki sets the external-LLM audit of Section N identifies as answerable (0.5% of the panel); the behavioral analyses elsewhere in the paper use the audited 750-question panel. The *decision position* p^* is the last prompt-token position: the next-token distribution at p^* produces the answer/abstain decision token, so p^* is exactly the position the decision-token loss L_{dec} (Section 5.3) supervises. At layer 0 the representation at p^* is the constant final template token's embedding, so layerwise decodability measures when information *arrives* at the decision position. The per-prompt *answer margin* is $a(x) = \mathrm{logsumexp}_{j\neq k}\, z(x)[j] - z(x)[k]$, where $z(x)$ are the logits at p^* and k is the first token of the abstention target; higher values favor answering. A layerwise logit lens applies the final normalization and unembedding to each layer's residual stream; at the final layer the lens output equals the model's own logits (verified in every sweep), and for earlier layers the standard logit-lens approximation caveat applies. Per checkpoint and layer, per-question-cell means of a are decomposed with the same ±1 coding as Equation (1) into the internal evidence-state margin $\hat{u}$, format margin $\hat{v}$, an intercept, and an interaction term; $\hat{u}, \hat{v}$ at the final layer are the headline quantities. Internal margins are logit-scale while the behavioral u, v are probit-rate-scale, so all cross-level comparisons are of sign, ordering, and time course. Layerwise linear probes decode two balanced targets — renderer chain-versus-raw and evidence sufficient-versus-insufficient (1,508/1,508 each; the content-fixed twins make the format target content-controlled) — under 5-fold question-id-grouped cross-validation with per-fold standardization; the primary probe is the hyperparameter-free mean-difference direction scored by out-of-fold AUROC, with an L_2-regularized logistic probe as a secondary check. Estimates carry question-id-cluster bootstrap 95% CIs. The measurements comprise 351 layer sweeps, each passing the decision-token, lens-agreement, and padding validation checks.

Availability at initialization. At LoRA initialization — the pretrained base reader, before any task-specific fine-tuning — the embedding layer carries no information at the decision position (AUROC 0.5 for both targets). After one attention block (layer 1) the format target is decodable at ceiling (AUROC 0.9999) versus 0.921 for sufficiency; the gap persists across the early band and widens toward the output end of the pretrained stack (layer 28: 1.000 versus 0.809). Averaged over layers 0–14, the format-minus-sufficiency AUROC gap is $+0.0271$ $[+0.0218, +0.0324]$ (Table A18a); the secondary logistic probe reproduces it in ceiling-compressed form ($+0.0007$ $[+0.0003, +0.0012]$). The depth at which each target first reaches out-of-fold AUROC 0.90 is identical (layer 1 for both), so the initialization advantage of format is a strength gap rather than an onset-depth gap — sufficiency is also linearly present throughout, only weaker and decaying toward the output head. As throughout (Section 3, Scope below), the evidence-state target bundles the constructed insufficiency with its item-count and length differences; a pretrained reader's layer-1 decodability of that target is expected

to draw on such coarse correlates rather than on multi-hop sufficiency verification, which bounds what the initialization advantage itself shows.

Formation trajectories. Table A17 tracks the final-layer margins along the $g{=}0.8$ trajectory at both seeds. The format margin $\hat{v}$ is persistently positive from step 4 of 249 — before the behavioral gate emerges between checkpoint steps 8 and 16 — while the evidence-state margin $\hat{u}$ becomes persistently positive at step 20; during the format burst (steps ≈4–12) $\hat{u}$ is pushed negative — with the 95% CI entirely below zero — to ≈ -0.48 at both seeds before recovering. Through the give-back phase $\hat{v}$ persists at high magnitude (≈1.6–2.6 across the retained checkpoints) with local fluctuations in both directions whose 95% CIs exclude zero (for example, seed 7 steps 48→64: -0.294 $[-0.315, -0.272]$; seed 17 steps 192→249: -0.534 $[-0.546, -0.523]$), and the point estimate of $\hat{u}$ crosses above $\hat{v}$ between steps 192 and 249 at both seeds (no interval test is reported for the crossing itself). The ordering contrasts, the give-back-window endpoint changes, and the fully-correlated-versus-balanced control contrasts are collected in Table A18a; the balanced arm retains a small residual format margin at the endpoint ($\hat{v} = +0.161$ $[+0.136, +0.185]$), while the paired fully-correlated-versus-balanced contrast is an order of magnitude larger ($\Delta\hat{v} = +9.356$ with $\Delta\hat{u} = -2.506$) — paralleling the correlated arm's lower behavioral sufficiency summary u.

Depth profile. At the trained endpoint ($g{=}0.8$, seed 7) the margin decomposition is near zero through layer 12, forms at layers ≈16–20 ($\hat{u}, \hat{v} \approx$ 0.6–1.8), and completes at the final layers ($\approx$ 2.2–2.5): decision-margin formation is a late-layer phenomenon in this reader, consistent with the late-layer placement of abstention overrides reported by Ling et al. (2025). Trained layerwise decodability shows mild late-layer compression (format AUROC falls from 1.00 in the early band to 0.88 at layer 20). The final-layer readout is exact by construction; earlier layers inherit the logit-lens approximation.

Dose tracking. The endpoint internal margins reproduce the behavioral dose patterns: $\hat{v}$ rises and $\hat{u}$ falls monotonically in the training-correlation dose g, at both seeds (Table A18b). Under the loss-reweighting intervention of Section 5.3, $\hat{u}$ rises monotonically in the L_{dec} dose (Spearman $+1.0$) and falls monotonically in the L_{rest} dose (-1.0) at both seeds, with internal contrast $\hat{D}_u = +1.457$ $[+1.386, +1.532]$ (seed 7) and $+2.424$ $[+2.330, +2.517]$ (seed 17); the earlier decision-loss dose ladder replicates internally with Spearman $+1.0$ at both full seeds, and the sparse seed-23 ladder preserves the ordering ($0.25\times$: $+1.471$; $1.0\times$: $+2.205$; $2.0\times$: $+2.384$).

Scope. The probe and lens reads are correlational instruments on one reader family and the development panel: they establish availability, formation order, and dose tracking of the internal margins, while the causal evidence stays at the loss level (Section 5.3). Format's near-ceiling layer-1 decodability is expected, since the rendering is an explicit scaffold-token lexical cue, so the initialization gap over the evidence-state target speaks to early feature *availability*; the evidence bearing on simplicity bias, or on competition between the rendering feature and the evidence-state feature, comes from the dose and loss interventions. The probe and lens target is the constructed evidence-state label, whose insufficient variant also differs in item count and length, so $\hat{u}$ is an internal evidence-state margin. The probe and lens measurements use the original evidence-state panels; the count- and length-matched evaluation of Section B assesses behaviour, and there the separations persist, so the behaviour the probes track is not a count or length artifact. Whether these internal directions carry the answer/abstain decision is a question for direction-level interventions (Section 7).

Table A17: Internal margin formation along the g=0.8 training trajectory: final-layer internal evidence-state margin $\hat{u}$ and format margin $\hat{v}$ at retained checkpoints, including every retained checkpoint of the give-back window, steps 48–249 (four-cell decomposition of the decision-position answer margin; protocol in Section J). The format margin is persistently positive from step 4 and the evidence-state margin from step 20, with a replicated transient suppression of $\hat{u}$ during the format burst; by the endpoint the point estimate of $\hat{u}$ rises past $\hat{v}$ at both seeds.

	Seed 7		Seed 17	
Step	$\hat{u}$	$\hat{v}$	$\hat{u}$	$\hat{v}$
0	−0.024	−0.167	−0.024	−0.167
4	−0.100	+0.119	−0.124	+0.090
8	−0.423	+0.525	−0.477	+0.497
10	−0.478	+0.814	−0.425	+0.560
16	−0.099	+0.936	−0.076	+1.206
20	+0.057	+1.272	+0.073	+1.686
48	+0.416	+2.188	+0.517	+1.818
64	+0.606	+1.894	+0.809	+2.621
96	+1.209	+2.264	+1.130	+1.624
128	+1.604	+1.996	+1.984	+2.172
192	+2.326	+2.350	+2.501	+2.604
249	+2.451	+2.212	+2.530	+2.070

Table A18: Internal margin contrasts and dose response (question-id-cluster bootstrap 95% CIs; protocol in Section J). **(a)** Primary and control contrasts: the availability gap at LoRA initialization (mean over layers 0–14 of the format-minus-sufficiency probe AUROC on the pretrained base reader, before task-specific fine-tuning), the give-back-window endpoint changes (g=0.8, seed 7, steps 48→249), the formation-ordering contrasts, the internal decision-loss-minus-complement contrast $\hat{D}_u$, and the paired fully-correlated-versus-balanced endpoint contrast with the balanced arm's own residual format margin. **(b)** Endpoint internal margins across the training-correlation dose g (step 249): $\hat{v}$ rises and $\hat{u}$ falls as g grows, at both seeds.

(a) Primary and control contrasts

Quantity	Estimate [95% CI]
Availability gap at initialization ($\text{AUROC}_F - \text{AUROC}_S$, layers 0–14)	+0.0271 [+0.0218, +0.0324]
Give-back window $\Delta\hat{u}$ (48→249)	+2.035 [+1.948, +2.123]
Give-back window $\Delta\hat{v}$ (48→249)	+0.024 [−0.020, +0.069]
$\hat{v}$ at step 16	+0.936 [+0.898, +0.974]
$\hat{u}(249) - \hat{u}(16)$	+2.550 [+2.444, +2.660]
$\hat{v}(g{=}0.8) - \hat{v}$(MARG) at step 249	+2.051 [+2.014, +2.089]
$\hat{D}_u$ ($L_{\text{dec}} - L_{\text{rest}}$ endpoint contrast), seed 7	+1.457 [+1.386, +1.532]
$\hat{D}_u$, seed 17	+2.424 [+2.330, +2.517]
A0 − MARG paired $\Delta\hat{v}$ at step 249	+9.356 [+9.271, +9.437]
A0 − MARG paired $\Delta\hat{u}$ at step 249	−2.506 [−2.611, −2.407]
MARG residual $\hat{v}$ at step 249	+0.161 [+0.136, +0.185]

(b) Endpoint dose response (step 249; seed 7 / seed 17)

Dose	$\hat{v}$	$\hat{u}$
g=0.5	+1.210 / +1.334	+3.005 / +2.924
g=0.8	+2.212 / +2.070	+2.451 / +2.530
g=0.9	+2.916 / +2.867	+1.644 / +1.635

K Vertical and Inference-Stage Detail

Native-label mapping. The substantive native labels — Entailment and Contradiction on ContractNLI, Yes and No on ShARC — map to *answer*. The native insufficiency labels map to *abstain*:

NotMentioned, meaning the contract does not support the relation, and Irrelevant, meaning the rule text cannot determine a decision. On a gold-unanswerable input a substantive output is therefore a native-label error — an answer the source documents do not support.

Vertical prompts and decision rules (verbatim). The ContractNLI units take each contract's annotated spans, in order, as the evidence items (identifier `S{i}`, sentence number i = span index, title `Contract`) and the hypothesis as the question block `Statement: {hypothesis}`; gold $\in$ {Entailment, Contradiction, NotMentioned}. The ShARC units take the non-empty lines of the rule-text snippet as the items (identifier `L{i}`, sentence number i = line index, title = the last path segment of the source URL) and the block `Question: {question}` followed, when present, by `Scenario: {scenario}` and one `Follow-up question: {q}` / `Follow-up answer: {a}` pair per dialogue turn; gold $\in$ {Yes, No, Irrelevant}. Both renderings apply the frozen chain and raw renderers of Section L to these items (e.g. `Step 3. [S12] Contract sentence 12: {text}` versus `Contract: {text}`), so within a unit the two prompts differ only inside the evidence body, which is checked for every unit. *ChatQA-1.5-8B* (Table 1a, rows 1–2, and the inference rules): the NVIDIA prompt frame `{system}\n\n{context}\n\nUser: {instruction} {question}\n\nAssistant:` with the system line `System: This is a chat between a user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions based on the context. The assistant should also indicate when the answer cannot be found in the context.`, the rendered evidence as the context, the dataset instruction — ContractNLI: `The context above consists of excerpts from a non-disclosure agreement. Classify the following statement against the contract. Answer with exactly one word: Entailment if the contract states or implies the statement, Contradiction if the contract states the opposite of the statement, or NotMentioned if the contract does not address the statement.`; ShARC: `Based only on the rule text above, answer the question with exactly one word: Yes, No, or Irrelevant. Answer Irrelevant if the question cannot be answered from the rule text. Take any scenario and follow-up answers into account.` — and the unit block as the question; exactly one beginning-of-sequence token is prepended by hand and the string is tokenized with `add_special_tokens=False`; prompts above $7{,}680$ input tokens are excluded; decoding is greedy with at most 256 new tokens. The decision comes from the frozen label parser, whose *answer* condition is a finite language: after the generation terminators are replaced by spaces, a response is an answer with label L if and only if it consists of letter-words separated by whitespace with at most one trailing period and no other character, its casefolded word sequence is one of 32 fixed declarative frames (the bare label; copula frames such as `it is`; and `the answer is`-style frames over `answer`/`classification`/`label`/`response` with optional `my`/`correct`/`final`) followed by the answer label L (`Entailment` or `Contradiction`; `Yes` or `No`) as its last word, and no line break stands immediately before the word `is`. `The answer is Entailment.` is an answer; `Entailment, not Contradiction.` is not: the parser refuses every response that adds prose, punctuation, quotation, a second label, a negator or a hedge to this language. Every other response is a non-answer and stays in the denominator; for the unparseable rate reported below it is routed, in order, as an abstention on an opening refusal core, on a withdrawal phrase anywhere, when the only non-negated unhedged label mention in the initial 30 tokens is the abstention label, or when no single label survives and a listed abstention phrase occurs in the normalized text; otherwise as unparseable (empty output, two or more distinct labels, only negated or hedged mentions, or a rejected answer-label commitment). The full frame enumeration, the terminator, refusal, withdrawal and abstention-phrase lists, and worked accept/reject examples are reproduced verbatim in the supplementary material; the answer condition as stated was checked against the frozen parser on every response of the four ChatQA-1.5-8B runs ($24{,}536$ rows, identical decisions) and on $4{,}608$ enumerated committing strings. The teacher-forced label scores of the inference rules append each candidate label to this same prompt. *TrustAlign-Qwen2.5-3B/7B* (Table 1a, rows 3–4): the checkpoint's own chat template (one user turn, generation prompt appended) around the official Trust-Align test-mode assembly `{instruction}\n\nQuestion: {question}\n\nDocument [1](Title: Evidence): {evidence}\n\nAnswer:`, where the instruction is the recipe's own: `Instruction: Write an accurate, engaging, and`

concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents. If none of the provided documents contain the answer, only respond with "I apologize, but I couldn't find an answer to your question in the search results.". Do not add further explanation as to why an answer cannot be provided, just state the response above as-is.; the rendered evidence is the single document, and the question is the unit block with one leading Question: label stripped (ShARC) or passed verbatim (Statement: ..., ContractNLI). Decoding is greedy with at most 300 new tokens (the official default), the output is post-processed as in the official evaluator (<|im_end|> and a trailing End. removed), and the decision is the official refusal rule: a fuzzy partial-ratio match of the normalized refusal flag I apologize, but I couldn't find an answer against the normalized output above the official threshold of 85 is an abstention, anything else an answer; an empty output would be rejected by the scorer rather than scored, and none occurred. *OCC-RAG-1.7B* (Table 1a, row 5, and the discordance rule): the checkpoint's own chat template with the rendered evidence passed as a single document, the unit block as the user message, the generation prompt appended and thinking disabled; decoding is greedy with at most $2{,}048$ new tokens (the model reasons before it answers); a unit is eligible only if, under both renderings, input $+\,2{,}048 \leq 32{,}768$ positions (no truncation). The decision is the model's native status field — exactly one <|status_start|>ANSWERABLE|UNANSWERABLE<|status_end|> — read as answer versus abstention; an output that reaches the cap without a parseable status counts as an abstention (a pre-registered amendment), an empty output and a missing or duplicated status field are recorded as non-answers, and the answer text between <|answer_start|> and <|answer_end|> is used only for descriptive statistics. The discordance rule's third pass asks the same checkpoint, in its native frame with the *raw*-rendered evidence: "Based only on the evidence provided, decide whether it contains enough information to answer the question below. Reply with exactly one word and nothing else: SUFFICIENT if the evidence is enough to answer, or INSUFFICIENT if it is not.", followed by the question. Only the reply's answer segment is parsed: a standalone (or leading-word) SUFFICIENT/INSUFFICIENT after stripping trailing punctuation decides; any other reply is unresolved and is scored as an abstention (a non-answer) in every reported statistic — this occurred on 21 of the 263 discordant items — and the native status field is never used as a fallback for this meta-question.

Output classification and parser audit. On the vertical benchmarks, every model output is routed by the deterministic parser (Section 3) into a substantive native label, a native abstention (the abstention label or a listed abstention phrase), or neither; the last class — including empty generations — is conservatively counted as a non-answer, so interface failures are never credited as substantive answers. This routing is a negligible fraction of the data: on the gold-unanswerable subsets, the unparseable rate is zero in five of the seven evaluated lanes (both TrustAlign checkpoints and OCC-RAG on ContractNLI, and both ChatQA ShARC splits) and at most 0.57% elsewhere (ChatQA on ContractNLI), with a chain$-$raw differential of at most 0.24pp — far below the reported contrasts. A stratified human audit of 210 sampled parser decisions on the ContractNLI and ShARC development splits (108 and 102 outputs) confirmed the parser's label mapping on every sampled output, including the routing of empty outputs to non-answers.

The ShARC answer-rate floor. All four released checkpoints were evaluated on both vertical benchmarks. ShARC is reported for ChatQA-1.5-8B alone because the other three sit at the answer-rate floor on the natively unanswerable Irrelevant subset: TrustAlign-Qwen2.5-7B answers no item under either rendering, TrustAlign-Qwen2.5-3B at most 2.5%, and OCC-RAG-1.7B 2.17% under both renderings, a contrast of exactly zero whose 95% interval (± 2.9pp) is wider than the answer rate itself. A contrast at the floor is indeterminate rather than null. ChatQA-1.5-8B instead answers 31.9% of Irrelevant items under chain and 23.2% under raw, far enough from the floor for an effect to be measurable. The same floor recurs for the frontier systems of Section F, which answer at most 0.4% of ShARC Irrelevant items, so this is a property of how readers treat that subset rather than of

any one checkpoint. ContractNLI shows no comparable floor for any reader evaluated here, which is why every checkpoint carries a ContractNLI row.

Margin decomposition and inference rules (referenced from Section 6.2). Writing an item's two per-rendering decision margins as $m_f(x) = h(x) + \delta(x)\, f$ with $f \in \{+1, -1\}$ separates a renderer-even common score $h(x)$ from a renderer-odd component $\delta(x)\, f$ (with $m_{\mathrm{c}} = m_{+1}$ and $m_{\mathrm{r}} = m_{-1}$); reading both renderings exposes the two parts directly. With reader weights frozen, each candidate label y is scored teacher-forced under each rendering by the *mean per-token* log-probability of its tokens appended to the prompt, $\ell(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p(y_t \mid \text{prompt}, y_{<t})$, where prompt and label are tokenized separately without special tokens and concatenated (the prompt already carries the model's own frame and one beginning-of-sequence token; Section K, vertical prompts); no whitespace is prepended to the label string, so each candidate label is scored from its no-leading-space token form, identically across labels and renderings, and the margin per rendering is $m = \max\{\ell(\text{Entailment}), \ell(\text{Contradiction})\} - \ell(\text{NotMentioned})$ on a ContractNLI development split; the pair is combined through rules that suppress the renderer-odd component. **AND** answers only when both renderings answer. **Margin fusion** answers when $S_\lambda = \frac{1}{2}(m_{\mathrm{c}} + m_{\mathrm{r}}) - \lambda |m_{\mathrm{c}} - m_{\mathrm{r}}| = h - 2\lambda|\delta|$ exceeds a threshold τ, interpolating from format cancellation (λ=0) through *maximin* (λ=$\frac{1}{2}$, i.e. $\min(m_{\mathrm{c}}, m_{\mathrm{r}})$) to stronger sensitivity penalties. Sweeping τ traces the operating curve of margin fusion and of the single-format margin baselines on the (unsupported-answer rate, correct-coverage) plane. **Discordance adjudication** invokes a third pass on disagreement; it is a discrete rule over categorical decisions and contributes a single operating point, not a τ-parameterized curve.

Correctness scoring. In the inference-rule correctness scoring, 824 of 864 method-level judgments were resolved by the deterministic matcher, and the remaining 40 (4.6%) were adjudicated by an external judge. Under this scoring, the in-sample margin-fusion reduction of Section 6.2 (ChatQA-1.5-8B) also holds relative to the binary chain, raw and AND rules.

Answer routing differs between rules. The binary chain, raw, AND and OR points use each rendering's greedy parsed label, whereas the fusion rules select the answer label by the argmax of the two renderings' averaged teacher-forced answer-label log-probabilities. The teacher-forced three-label argmax agrees with the greedy label on 81.4% of readings, and re-routing AND's 202 answered sufficient items through the fused router changes 9 labels (correct coverage 0.341 greedy versus 0.329 re-routed); the comparison below therefore gives every rule its own teacher-forced router over the two answer labels.

Single-format thresholds and cross-fitted selection. To ask whether the fusion gain comes from paired-format information, score fusion or threshold selection, we sweep a threshold τ on each single-rendering margin alone (answer iff $m > \tau$; label from that rendering's answer-label log-probabilities) and on the fusion scores, over the grid of every distinct score value among the calibration units plus its two endpoints. Thresholds are chosen on calibration documents and read on held-out documents: the 58 ContractNLI development documents are assigned once to five folds (documents sorted by id, permuted with seed 20260907, round-robin); on each calibration set τ is the largest grid value whose correct coverage reaches the target c^*=0.341 (AND's correct coverage) or the smallest whose unsupported-answer rate is at most $u^* \in \{0.05, 0.08, 0.114\}$; the better single-format rule is chosen on the calibration folds only; the five held-out folds are pooled. Intervals are document-level bootstraps (B=10,000) that repeat fold assignment, calibration and selection in every replicate. At c^*=0.341 the raw-margin threshold is the better single-format rule in every fold and achieves 0.083 unsupported at 0.347 correct coverage; λ=0 fusion achieves 0.090 at 0.347, a paired difference of $+0.007$ $[-0.021, +0.033]$ in unsupported rate and $+0.000$ $[-0.016, +0.027]$ in correct coverage — no resolved advantage in either direction; the chain-margin threshold reaches 0.143 at 0.343. At the unsupported-rate targets the correct-coverage differences between fusion and the raw threshold are -0.005 $[-0.043, +0.029]$, -0.007 $[-0.034, +0.029]$ and -0.016 $[-0.037, +0.016]$; maximin (λ=$\frac{1}{2}$) is likewise unresolved, and λ=1 loses coverage (-0.044 $[-0.085, -0.007]$ at u^*=0.114). Holding the chain router fixed and changing only the gate score gives no resolved difference (-0.029 $[-0.074, +0.012]$ unsupported at c^*). Holding the fusion gate fixed and changing the router from chain to fused raises correct coverage by $+0.010$ $[+0.002, +0.022]$, $+0.016$ $[+0.007, +0.026]$ and $+0.017$ $[+0.009, +0.029]$ at the three unsupported-rate targets; the same comparison against the raw router gives 0.000 $[-0.009, +0.003]$, -0.005 $[-0.014, +0.003]$ and -0.009 $[-0.019, +0.002]$

(unsupported rates identical by construction), so averaging the two label scores improves on the chain-only router, while its difference from the raw router is unresolved. The in-sample λ=0 curve read at AND's coverage reproduces the 0.080 of Section 6.2. These comparisons all use the ContractNLI development documents with cross-fitted selection; Figure A4b overlays the single-format curves and Table A19a lists the cross-fitted points.

Selection and bootstrap rules (deterministic). Decisions are *answer iff score* $> \tau$ (equal scores share a decision). In each fold the candidate set is $\{-\infty\} \cup \{\text{distinct calibration scores}\} \cup \{+\infty\}$; for an unsupported-rate target u^*, τ^* is the smallest candidate whose calibration unsupported rate is $\leq u^*$ ($+\infty$ always qualifies); for a coverage target c^*, τ^* is the largest candidate whose calibration correct coverage is $\geq c^*$, and if none qualifies $\tau^* = -\infty$ (answer on all) and the target is flagged unattainable for that fold. The better single-format rule is the one with the lower calibration unsupported rate at a coverage target (higher calibration coverage at an unsupported-rate target), ties broken by the other coordinate and then in favour of the chain margin; λ is never selected. Held-out decisions are pooled over the five folds and both achieved coordinates are reported. Bootstrap: B=10,000 draws of the 58 documents with replacement (`numpy` RandomState 20260907); a document drawn k times keeps its identity with multiplicity k and all its copies sit in one fold; folds are reassigned over the distinct drawn documents by a fresh permutation, and calibration, selection and evaluation are repeated in full; intervals are percentile 2.5/97.5 of the paired differences. A secondary conditional interval keeps the full-data rules fixed and resamples evaluation documents the same way.

Cost of the second reading. Label scoring forwards the entire prompt once per candidate label, three labels per rendering, with no prompt reuse: a single-format rule processes about $3 \times 3{,}259 \approx 9.8$k (chain) or $3 \times 2{,}428 \approx 7.3$k (raw) prompt-token positions per item over three forwards, and fusion about 17.1k over six; unique prompt lengths are 3.3k, 2.4k and 5.7k tokens. OCC-RAG's discordance adjudication needs two generations plus a third on disagreement. Wall-clock latency was not measured.

L Setup Detail

Renderers. The chain and raw renderers (Section 3) serialize the same deterministically selected, ordered evidence items. For the fleet readout of Figure 1 and for the core arms, chain joins the items with newlines, item i rendered as

```
Step {i}.  [{evidence_id}] {title} sentence {sent_id}:  {text}
```

and raw renders each item as `{title}:  {text}` joined by blank lines; items and order are identical, so only the scaffold tokens differ, and the chain scaffold adds both wording and length (the matched-length control of Section 4.3, Section A, tests the added-length ingredient against the scaffolded renderings). The core arms' training rows and their 750-question readouts use these two renderers on content-fixed twins; the matched-length control (Section A), the unseen-renderer family (Section C), the marker family of the corpus intervention (Section D) and the vertical ContractNLI/ShARC prompts (Section K, which lists them verbatim) define their own serializations where they differ. Recipe A's training positives are the stored chain rendering of each HotpotQA package's retrieved evidence chain, and its negatives are three same-question distractor paragraphs from the package's raw-context rendering — each `[P{idx}] {title}: {sentences}`, blank-line separated, the top three by lexical overlap with the question among gold-free paragraphs, emitted with probability 0.82 per question and never for yes/no questions or questions with fewer than three gold-free paragraphs; Recipe B renders its positives (the full chain) and negatives (the gold-removed insufficient subset, up to three seeded-uniform items) with the two renderers above. The fleet readout prompt, held fixed across the adapter-on and adapter-off legs, is `Answer the question using only the evidence below. Give a short answer.\n\nQuestion: {q}\n\nEvidence:\n{e}\n\nAnswer:`; Recipe B trains with `Read the evidence and answer the question. Use only the evidence provided. Give a short answer.` followed by the same question/evidence/answer scaffold. Figure A1 shows a real behavioral consequence on one development row.

Table A19: Inference-rule operating points on the ContractNLI development split (frozen reader weights). **(a)** ChatQA-1.5-8B on the (unsupported-answer rate, correct-coverage) plane: the binary AND point and the λ=0 fusion curve read in-sample at AND's coordinates, and the cross-fitted operating points (thresholds chosen among calibration-document scores, read on held-out documents, five document folds) at the common calibration target c^*=0.341, with the paired fusion−raw-threshold difference and its repeated-selection document-bootstrap 95% CI (Section K). **(b)** OCC-RAG-1.7B discordance-adjudication paired contrasts (adjudication − baseline; paired qid-cluster bootstrap 95% CIs).

(a) ChatQA-1.5-8B operating points

Rule	Unsupported-answer rate	Correct coverage
Binary AND	0.114	0.341
λ=0 (at AND's coverage)	0.080	0.341
λ=0 (at AND's unsupported-answer rate)	0.114	0.364
Cross-fitted at c^*=0.341 (achieved rates)		
Chain-margin threshold	0.143	0.343
Raw-margin threshold (best single-format)	0.083	0.347
λ=0 fusion	0.090	0.347
Maximin (λ=0.5)	0.097	0.343
λ=0 − raw threshold [95% CI]	$+0.007$ $[-0.021, +0.033]$	$+0.000$ $[-0.016, +0.027]$

(b) OCC-RAG-1.7B adjudication contrasts

Baseline	Metric	Adjudication − baseline [95% CI]
Chain-only	unsupported-answer rate (unanswerable)	-7.04pp $[-10.60, -3.65]$
Chain-only	correct coverage	$+1.05$pp $[-1.64, +3.49]$
Binary AND	correct coverage	$+8.19$pp $[+5.82, +10.62]$
Binary AND	unsupported-answer rate (unanswerable)	$+4.85$pp $[+3.10, +6.77]$

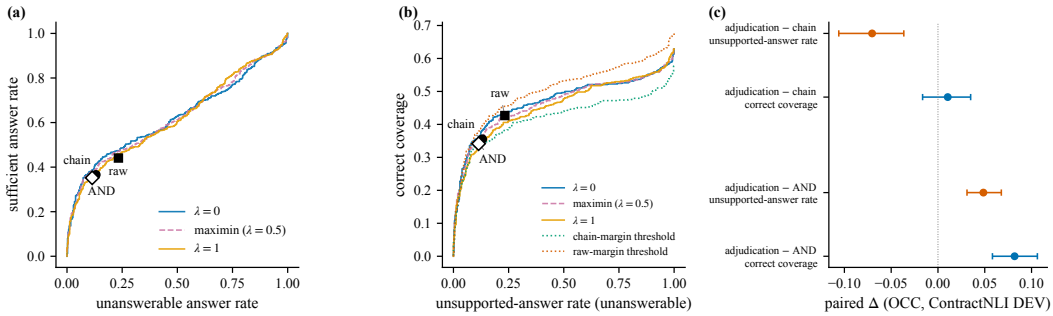

Figure A4: Inference-only rules on frozen reader weights (ContractNLI development split). **(a)** ChatQA-1.5-8B unscored answer-rate plane: margin-fusion curves ($\lambda \in \{0, 1\}$ and maximin λ=0.5) and binary chain/raw/AND points. **(b)** ChatQA-1.5-8B scored plane (unsupported-answer rate vs correct coverage): the λ=0 and maximin curves pass to the upper left of the binary AND point (lower unsupported rate at fixed coverage, higher coverage at fixed unsupported rate), an improvement over the binary rules (Table A19a); the dotted single-format threshold curves (chain margin, raw margin; in-sample sweeps with each rendering's own answer-label router) show the raw-margin sweep running alongside the fusion curves (Section K). **(c)** OCC-RAG-1.7B paired contrasts of the adjudication rule versus chain and versus AND; values in Table A19b.

Readers and training. The nine-reader gate fleet are instruction-, chat-, or IT-tuned checkpoints — Qwen2.5-0.5B/1.5B/7B, Llama-3.2-1B/3B, Gemma-2-2B, Phi-3.5-mini, Yi-1.5-6B, SmolLM2-1.7B — each with the recipe's LoRA adapters. The core training-side readers are the *base* checkpoints Llama-3.2-3B (Grattafiori et al., 2024; Meta AI, 2024) and Qwen3-8B-Base (Yang et al., 2025), with OLMo-2-7B (Team OLMo et al., 2025) in the multi-seed extension. The external and inference lanes use off-the-shelf released checkpoints (ChatQA, TrustAlign, OCC-RAG, and the additional

benchmark checkpoints of Table A5), evaluated without retraining; only the core intervention set are base models. LoRA rank 16; the core arms train for $T{=}249$ steps with checkpoint readouts at steps 48 and 249; decoding, parsing, scoring and the per-recipe hyperparameters (Table A20) follow; seeds as stated per result.

Decoding, parsing, and scoring. The fleet readout decodes greedily (`do_sample=False`, at most 64 new tokens); the core-arm 750-question readout decodes greedily with one beam and at most 128 new tokens, stopping on the union of the model's and the tokenizer's end-of-sequence ids. The parser takes the `answer` field of a JSON object if the output contains one, salvages a truncated JSON answer otherwise, and else takes the initial non-degenerate line after stripping answer prefixes; an empty or degenerate output parses to the empty string. The abstention predicate holds when the normalized parsed answer contains any of "insufficient evidence", "not enough evidence", "cannot determine", "can't determine", "not provided" or "unknown"; an output is scored as an answer otherwise, so an empty output or a refusal phrased outside this list counts as an answer (the supplementary per-series table counts these cases).

Training arms. Table A22 summarizes the training arms of Section 3: the dose ladder, the balanced recipes, and the matched reversal. The reversal swaps only the renderer→target assignment, holding content, question ids, target multiset, and mixture fixed. Adapters are trained with cross-entropy unless noted; the balanced recipes are additionally trained with GroupDRO (Sagawa et al., 2020a) (the -DRO arms). The main training seed is 7, with seeds 17/23 in the multi-seed extension of Section H.

Training recipes and hyperparameters. Table A20 lists the original fleet and core-arm configurations with the provenance of each value; the additional native-template check is described below. Values not set explicitly in the training code take the defaults of the package versions recorded in the run manifests, which are the versions that ran, not current defaults. The core-arm manifests record the learning rate, effective batch, sequence length, step budget, row visits, LoRA configuration and package versions; their optimizer, schedule, warm-up, weight decay and gradient clipping are not set in the training code at the recorded commits and are therefore the recorded version's defaults. The core arms train on the V3 prompt (byte-identical to the fleet readout prompt) with the gold answer string as the answer target and "cannot determine" as the single abstention target; for the configurations in Table A20, the training sequence is `prompt + " " + target + EOS` under the tokenizer's default special tokens and no chat template, truncated and padded to $1{,}024$ tokens, with the loss masked on the prompt span (its length measured by tokenizing the prompt alone) and on padding, so only the target and its EOS are supervised.

Matched training budget. Every training arm — A0, the dose arms, MARG, and CROSS, across readers and seeds — runs under an identical budget: 7,940 training row-visits over 249 optimization steps at effective batch size 32 — 248 full steps of 32 row-visits and a final step carrying the residual 4 ($248 \times 32 + 4 = 7{,}940$; the recorded run manifests confirm 249 steps at 7,940 visits per arm) — with identical question-id exposure and the same LoRA configuration, optimizer, decoding, and parser. Sequences are padded to a fixed length of 1024, so the *per-run* token budget is identical across arms ($7{,}940 \times 1024 = 8{,}130{,}560$ positions; each full step processes 32×1024). The exact-crossing recipe pairs both renderings within items: it draws its 7,940 visits from 7,940 distinct serialized rows, whereas A0 and the marginal recipes visit 3,970 rows twice — this distinct renderer exposure is the intervention itself. Re-tokenizing the training rows with each reader's own tokenizer (Table A21) shows that this changes the realized token budget only marginally: within each reader the supervised target-token totals are identical across arms, and the non-padding input-token (content) totals differ by less than 0.18% (CROSS-CE versus A0: 0.012% Llama-3.2-3B, 0.004% Qwen3-8B). The recipes are therefore matched on optimization steps, question–evidence-state presentations, padded per-step token budget, and target tokens, and near-matched (below 0.18%) on content tokens.

Nine-reader fleet: adapter series, recipes, seeds, and raw rates. The fleet comprises seven recipe–seed configurations in four groups: the deterministic Recipe-A re-runs, an earlier single-run Recipe-A series, and Recipe B forward and reversed. The complete per-series readings — raw frozen-parser answer rates (the historical reading) in every cell, the adapter DiDs with their intervals, the empty-corrected reading, per-cell empty and refusal-like counts, and the identity of each series — are provided for all seven configurations as a machine-readable table in the supplementary material;

Table A20: Training recipes and hyperparameters. Provenance of each value: E, set explicitly in the training code or recorded in the run manifest; D, default of the recorded package version (transformers 5.12.1, peft 0.19.1, torch 2.5.1 for every manifest-bearing run). The historical fleet adapters carry an adapter configuration but no run manifest; their column repeats the recipe they were trained with.

	Fleet Recipe A, historical	Fleet Recipe A, re-runs	Fleet Recipe B (forward / reversed)	Core arms (A0, dose, MARG, CROSS)
readers	nine 0.5–7B fleet	same	same	Llama-3.2-3B, Qwen3-8B; OLMo-2-7B (multi-seed)
seeds	7	7 / 17 / 23 (E)	7 / 17; reversal 7 (E)	7; 17 / 23 in the multi-seed extension (E)
prompt; targets	V3 prompt; gold / "insufficient evidence"	same (E)	Recipe-B prompt; gold / "cannot determine" (E)	V3 prompt (identical) on content-fixed twins; gold / "cannot determine" (E)
rows	as the seed-7 re-run	1,000 qids; 1,725 / 1,741 / 1,735 rows at seeds 7 / 17 / 23 (E)	447 qids, 894 rows, 50/50 (E)	3,970 rows visited twice; 7,940 distinct rows for CROSS (E)
epochs / steps	2 epochs	2 epochs (E)	2 epochs (E)	249 steps over 7,940 row-visits (E)
effective batch	32	32: per-device ladder $\times$ accumulation (E)	same (E)	32: per-device 2 $\times$ accumulation 16 (E)
learning rate	10^{-4}	10^{-4} (E)	10^{-4} (E)	10^{-4} (E)
optimizer	AdamW, β=$(0.9, 0.999)$, ϵ=10^{-8} (D)	same (D)	same (D)	same (D)
schedule; warm-up	linear decay to zero; none (D)	same (D)	same (D)	same (D)
weight decay; clip	0; 1.0 (D)	same (D)	same (D)	same (D)
sequence; loss	prompt + space + target + EOS, tokenizer-default special tokens, no chat template, truncated/padded to 1,024; prompt span and pad positions masked (-100), target + EOS supervised	same (E)	same (E)	same (E)
precision; length	bf16; 1,024	bf16; 1,024 (E)	bf16; 1,024 (E)	bf16; 1,024 (E)
LoRA	r=16, α=32, dropout 0.05, q/k/v/o (E, adapter configuration)	same (E)	same (E)	same (E)
determinism	not recorded	strict, eager attention (E)	strict, eager attention (E)	strict (E)
versions	peft 0.19.1 (E, adapter configuration)	transformers 5.12.1, peft 0.19.1, torch 2.5.1 (E)	same (E)	same; python 3.10 (E)

Table A21: Token-budget audit for the seed-7 recipes: non-padding input (content) tokens and supervised target tokens, summed over the realized 7,940-visit schedule and recomputed from the training rows with each reader's own tokenizer. Target totals are identical across arms; non-padding input totals differ by less than 0.18% within each reader (maximum pairwise spread 0.171% Llama, 0.164% Qwen).

Reader	Recipe	Input tokens	Target tokens
Llama-3.2-3B	A0	2,254,218	29,420
Llama-3.2-3B	MARG-CE	2,258,064	29,420
Llama-3.2-3B	CROSS-CE	2,254,494	29,420
Qwen3-8B	A0	2,345,782	30,822
Qwen3-8B	MARG-CE	2,349,620	30,822
Qwen3-8B	CROSS-CE	2,345,886	30,822

Table A23 carries the semantic reading of the six re-run and Recipe-B configurations (the historical series has the parser readings only). Figure 1a draws the seed-7 deterministic Recipe-A re-run under the semantic reading; across seeds 7/17/23 the re-runs give an insufficient-stratum gate with the per-run interval above zero at every seed for eight of the nine readers (the smallest, Phi-3.5-mini's, $+0.576$ / $+0.464$ / $+0.540$), the exception being Qwen2.5-7B ($+0.028$ / $+0.168$ / $+0.016$, unresolved at seeds 7 and 23); under the frozen parser the same eight readers are positive at every seed and Qwen2.5-7B is again the exception ($+0.048$ / $+0.200$ / $+0.044$). The earlier single-run series of the same recipe at seed 7 — adapters carrying an adapter configuration (LoRA r=16, α=32, dropout

Table A22: Training arms. The dose g sets the renderer–target correlation ($g{=}1$ fully correlated, $g{=}0$ marginal); $\hat{\rho}$ is the realized counterexample share. CROSS-CE and the reversal are content-fixed constructions rather than dose points. Each recipe is trained with cross-entropy (the "-CE" arms) or GroupDRO (the "-DRO" arms).

Arm	Construction	$\hat{\rho}$
A0	chain→answer, raw→abstain (dose $g{=}1$)	0%
$g{=}0.5$–0.95	renderer drawn per item at dose $g{\in}[0.5, 0.95]$	24.01%–2.52%
MARG	renderer by a per-row fair coin (dose $g{=}0$)	49.37%
CROSS-CE	both renderings paired within each item	$\frac{1}{2}$
reversal	swaps only the renderer→target assignment	—

0.05 on q/k/v/o) but no run manifest, their deterministic-training settings not recorded (Table A20) — has a larger pooled gate ($+0.949$ $[0.940, 0.959]$ sufficient, $+0.906$ $[0.890, 0.920]$ insufficient), and its Qwen2.5-7B estimate of $+1.032$ is the one the re-runs do not reproduce. The pooled intervals throughout are question-bootstrap intervals conditional on these nine readers and these fixed training seeds. Recipe B (Figure 1b) is an independent format-correlated recipe — its own prompt, the abstain target "cannot determine", 447 questions and 894 rows, negatives rendered from gold-removed insufficient item subsets — whose forward gates are smaller than Recipe A's and whose matched reversal swaps only the renderer→target assignment; the reversal holds for 8/9 readers under the semantic reading and under the frozen parser alike, and Phi-3.5-mini misses the positive-to-negative criterion under either because its forward effect is not positive (negative under the semantic reading, unresolved under the parser, positive only under the parser's empty-corrected variant; Table A23 and the supplementary per-series table). Because the frozen-parser statistic scores an empty output or an off-list refusal as an answer (Section L), the supplementary per-series table also reports, per adapter-on insufficient stratum, the number of empty outputs and of outputs matching an off-list refusal expression — a descriptive screen applied to the full output string in the precedence frozen abstain, empty, refusal-like, other; the expression is one fixed case-insensitive pattern over negated-verb, "no information"-style and insufficiency phrasings, reproduced verbatim in the supplementary material. The screen itself is a surface-pattern match that makes no semantic reading of the outputs it classifies; the semantic reading is the third layer below, whose judged results are in Table A23. For Phi-3.5-mini's Recipe-B forward adapter on insufficient items, for instance, the chain cell holds 1 frozen abstention, 2 empty outputs, 62 refusal-like and 185 other outputs, and the raw cell 0, 21, 50 and 179.

Three readings of "answer". The fleet readouts were scored by a frozen parser: `parse_model_answer_v2` extracts the answer string and the abstention predicate marks it as an abstention iff the *normalized* answer (casefolded, punctuation and articles removed, whitespace collapsed) contains one of six listed phrases. Two consequences follow. A whitespace-only response, and a refusal worded outside the list, both score as answers. And the listed literals are matched against the normalized text while themselves un-normalized, so the listed "can't determine" can never occur in a normalized string and is inert: five of the six phrases are live, and an output whose normalized text reads "cant determine" scores as an answer. Across all 63 series that case arises 35 times, all in one reader's adapter-off legs. We therefore report every re-run and Recipe-B series under three readings, the semantic one being the primary; the historical series carries the two parser readings. **(L1)** the frozen-parser non-abstention rate, the historical reading, unchanged; its three-seed pooled Recipe-A re-run DiDs are $+0.802$ (sufficient) and $+0.780$ (insufficient), the values quoted in Section 4.1. **(L2)** an *empty-corrected* reading, in which a whitespace-only response is a non-answer and the denominator stays at 250; it is applied identically to every cell and both legs of all 63 series (supplementary per-series table), and nine series move by at least 0.05 in the insufficient stratum — Phi-3.5-mini's two Recipe-B forward series, whose DiD moves from unresolved to positive, and SmolLM2-1.7B's seven, whose adapter-off cells are almost entirely empty. **(L3)** a *semantic answer-commitment* reading, in which a blinded judge labels each response COMMIT, ABSTAIN or INVALID and only COMMIT counts as an answer; it is the reading behind Figure 1 and the main text. Phi-3.5-mini's Recipe-B estimates referenced from Section 5.1 are forward -0.104 $[-0.164, -0.044]$ and reversed -0.140 $[-0.208, -0.076]$ — both negative, so the positive-to-negative reversal criterion is not met.

SmolLM2 native-template confirmation. SmolLM2-1.7B produced predominantly empty adapter-off outputs under the shared plain-text template. We therefore additionally trained and evaluated it with its native chat template using Recipe B at seeds 7, 17 and 23, with the Recipe-B instruction as the system turn and the question and evidence as the user turn. On the same audited 250-question panel, the native baseline produced no empty outputs ($0/500$ in each evidence state). Forward and reversed tuning yielded positive ($+11.2$ to $+11.6$ pp) and negative (-42.4 to -32.0 pp) insufficient-evidence chain-minus-raw semantic answer-rate gaps, respectively, with every per-seed 95% question-bootstrap interval excluding zero.

The semantic reading: instrument, judge, coverage and validation. The judge is `gpt-5.6-luna` at reasoning effort `max`, a single annotator run tool-free through its command-line client, one fresh session per batch. It sees only ten items at a time as `QUESTION:` / `RESPONSE:` pairs under opaque identifiers — never the reader, series, leg, rendering, evidence or gold answer — with no question repeated inside a batch and every stratum's items spread evenly over the ten batch positions. The decision rules are the frozen ontology of Section F used unchanged, plus a three-point clarification amendment — label frequencies carry no expectation, and a set of almost entirely empty responses is labelled UNCLEAR when that is what the text is; when a response runs on past its answer into a question the system invented for itself, only the stance toward the given question is judged; a bare short answer with no sentence around it is still a commitment — with two further calibration examples, reproduced verbatim in the supplementary material; a batch is accepted only if the reply is exactly the requested JSON with one label per listed identifier, the session shows no tool use, and its record confirms the judge model. Every planned item carries a valid label before any statistic is computed. Coverage is every response of the fleet readouts the main text rests on — the 27 Recipe-A series in all four cells, the 27 Recipe-B series in the insufficient stratum, and each reader's shared adapter-off leg in all four cells, $49{,}500$ responses — plus $4{,}800$ responses of the core arms on the primary 750-question panel (100 questions each for A0, MARG-CE and CROSS-CE and 50 for each dose-ladder arm, all four cells; the question lists are recorded with the judged sample): $54{,}300$ in all, every one judged under the same instrument. Identical question–response pairs judged more than once received identical labels for 95.7% of $5{,}049$ such pairs ($[95.1, 96.3]$). As a check on the single judge, a second model (`gpt-5.6-terra`, effort `max`, same instruction and channel) re-labelled a stratified sample of 750 judged responses blind to the other judge's labels: agreement 98.5% $[97.4, 99.2]$, Cohen's $\kappa = 0.97$ on the three labels and 0.97 on the commit-versus-not decision, and no sampling stratum (six groups of 125: Recipe A on, sufficient and insufficient; Recipe B on, insufficient; adapter off, sufficient and insufficient; core arms) below $\kappa = 0.94$ — an agreement check between two model judges, not a measure of accuracy against a human standard; no label was changed on its account. Table A23 reports the semantic DiDs of every re-run and Recipe-B series, Table A24 the parser-versus-judge disagreement by reader and leg, and Table A25 the core-arm subset, on which sufficiency-oriented accuracy under the two readings differs by at most 0.010 for every arm. In the original fleet readouts, the disagreement is concentrated on the adapter-off legs, where base readers emit prose that the parser scores as an answer and the judge does not (31.4% of non-empty insufficient-stratum responses, against 3.4% on the Recipe-A adapter-on legs). Two limits remain: the reading is that of one judge, validated by a second on a sample rather than by a three-annotator majority as in Section F, and blinding the metadata does not fully blind the rendering, because an adapter-off response that runs on into text it generates itself can echo the scaffold.

Semantic re-scoring of the mechanism analyses. The three parser-scored mechanism analyses of the body — the seven-arm dose ladder (both readers, seed 7), the decision-loss control (Llama-3.2-3B, seeds 7 and 17) and the seven give-back cells (steps 48 and 249) — were re-judged in full on the audited 750-question panel, all four cells ($138{,}000$ responses), by the same blinded judge as the fleet reading, reusing a previously assigned label only where an identical question–response pair had been judged unanimously and judging every other distinct pair once; the coefficients were then recomputed with the body's own estimator, changing only the answer label. Every registered trend holds under the semantic reading: the format effect rises with the training correlation for both readers, the evidence-state sensitivity declines from MARG to $g{=}0.95$, the Spearman trends and the crossover arms are unchanged, every SOA contrast to MARG stays negative under the joint simultaneous bands, the decision-loss ladder stays monotone with D_u positive at both seeds, and all seven give-back cells keep $\Delta U > 0$ and $\Delta V < 0$ (Table A26). Magnitudes move: v_{policy} is lower under the semantic reading at every interior arm (by up to 0.127 at Llama-3.2-3B $g{=}0.95$), the slope and Δu changes are resolved under the paired bootstrap, four give-back contrasts shift with intervals excluding zero, and

Table A23: Adapter DiDs under the semantic answer-commitment reading (L3): 250 questions per cell, paired qid-bootstrap 95% intervals; $^+$ / $^-$ mark intervals entirely above / below zero. Every position of every listed series is judged (adapter on and off, both renderings); readers are ordered by parameter count. **(a)** Recipe-A deterministic re-runs, insufficient stratum, with the frozen-parser (L1) value for the seed-7 re-run; the L1 estimates and intervals are reproduced from the supplementary per-series table, which carries L1 and L2 for all series. **(b)** Recipe-B forward (seeds 7 and 17) and matched-reversal (seed 7) series, insufficient stratum. **(c)** Recipe-A re-runs, sufficient stratum. Pooled rows are the equal-weight nine-reader mean under a joint question bootstrap.

(a) Recipe A, insufficient stratum

Reader	A s7 (L1)	A s7 (L3)	A s17 (L3)	A s23 (L3)
Qwen2.5-0.5B	$+0.920\,[0.876, 0.964]^+$	$+0.912\,[0.840, 0.980]^+$	$+0.924\,[0.852, 0.992]^+$	$+0.892\,[0.812, 0.968]^+$
Llama-3.2-1B	$+0.740\,[0.684, 0.792]^+$	$+0.688\,[0.592, 0.780]^+$	$+0.712\,[0.628, 0.796]^+$	$+0.664\,[0.580, 0.748]^+$
Qwen2.5-1.5B	$+0.972\,[0.932, 1.012]^+$	$+0.860\,[0.800, 0.920]^+$	$+0.872\,[0.812, 0.932]^+$	$+0.868\,[0.808, 0.928]^+$
SmolLM2-1.7B	$+0.668\,[0.608, 0.724]^+$	$+0.560\,[0.488, 0.632]^+$	$+0.728\,[0.668, 0.784]^+$	$+0.564\,[0.492, 0.632]^+$
Gemma-2-2B	$+0.828\,[0.776, 0.880]^+$	$+0.772\,[0.704, 0.832]^+$	$+0.868\,[0.812, 0.920]^+$	$+0.832\,[0.768, 0.892]^+$
Llama-3.2-3B	$+0.904\,[0.860, 0.944]^+$	$+0.852\,[0.780, 0.920]^+$	$+0.892\,[0.824, 0.956]^+$	$+0.904\,[0.840, 0.968]^+$
Phi-3.5-mini	$+0.796\,[0.744, 0.844]^+$	$+0.576\,[0.496, 0.652]^+$	$+0.464\,[0.384, 0.544]^+$	$+0.540\,[0.464, 0.616]^+$
Yi-1.5-6B	$+0.980\,[0.960, 0.996]^+$	$+0.932\,[0.876, 0.988]^+$	$+0.940\,[0.884, 0.996]^+$	$+0.944\,[0.888, 1.000]^+$
Qwen2.5-7B	$+0.048\,[-0.004, 0.104]$	$+0.028\,[-0.024, 0.080]$	$+0.168\,[0.096, 0.240]^+$	$+0.016\,[-0.040, 0.072]$
pooled	$+0.762\,[0.744, 0.778]^+$	$+0.687\,[0.664, 0.709]^+$	$+0.730\,[0.708, 0.752]^+$	$+0.692\,[0.669, 0.714]^+$

(b) Recipe B, insufficient stratum

Reader	B fwd s7 (L3)	B fwd s17 (L3)	B rev s7 (L3)
Qwen2.5-0.5B	$+0.360\,[0.276, 0.444]^+$	$+0.364\,[0.280, 0.448]^+$	$-0.636\,[-0.712, -0.556]^-$
Llama-3.2-1B	$+0.244\,[0.160, 0.332]^+$	$+0.228\,[0.140, 0.320]^+$	$-0.316\,[-0.400, -0.236]^-$
Qwen2.5-1.5B	$+0.132\,[0.056, 0.208]^+$	$+0.084\,[0.012, 0.156]^+$	$-0.272\,[-0.344, -0.200]^-$
SmolLM2-1.7B	$+0.432\,[0.356, 0.504]^+$	$+0.496\,[0.424, 0.564]^+$	$-0.660\,[-0.724, -0.596]^-$
Gemma-2-2B	$+0.096\,[0.040, 0.148]^+$	$+0.220\,[0.156, 0.284]^+$	$-0.592\,[-0.664, -0.520]^-$
Llama-3.2-3B	$+0.440\,[0.360, 0.516]^+$	$+0.540\,[0.460, 0.620]^+$	$-0.596\,[-0.676, -0.516]^-$
Phi-3.5-mini	$-0.104\,[-0.164, -0.044]^-$	$-0.092\,[-0.156, -0.028]^-$	$-0.140\,[-0.208, -0.076]^-$
Yi-1.5-6B	$+0.584\,[0.508, 0.660]^+$	$+0.620\,[0.540, 0.700]^+$	$-0.752\,[-0.832, -0.672]^-$
Qwen2.5-7B	$+0.200\,[0.124, 0.272]^+$	$+0.148\,[0.080, 0.216]^+$	$-0.316\,[-0.396, -0.236]^-$
pooled	$+0.265\,[0.237, 0.292]^+$	$+0.290\,[0.261, 0.319]^+$	$-0.476\,[-0.502, -0.449]^-$

(c) Recipe A, sufficient stratum

Reader	A s7 (L3)	A s17 (L3)	A s23 (L3)
Qwen2.5-0.5B	$+1.012\,[0.956, 1.068]^+$	$+0.980\,[0.920, 1.036]^+$	$+0.992\,[0.936, 1.048]^+$
Llama-3.2-1B	$+0.936\,[0.880, 0.992]^+$	$+0.924\,[0.868, 0.980]^+$	$+0.900\,[0.836, 0.964]^+$
Qwen2.5-1.5B	$+1.008\,[0.972, 1.044]^+$	$+0.980\,[0.940, 1.020]^+$	$+1.008\,[0.972, 1.044]^+$
SmolLM2-1.7B	$+0.796\,[0.736, 0.848]^+$	$+0.916\,[0.872, 0.952]^+$	$+0.820\,[0.764, 0.872]^+$
Gemma-2-2B	$+0.844\,[0.788, 0.900]^+$	$+0.880\,[0.832, 0.928]^+$	$+0.892\,[0.844, 0.940]^+$
Llama-3.2-3B	$+0.784\,[0.716, 0.852]^+$	$+0.848\,[0.784, 0.912]^+$	$+0.884\,[0.824, 0.944]^+$
Phi-3.5-mini	$+0.548\,[0.480, 0.616]^+$	$+0.432\,[0.368, 0.496]^+$	$+0.500\,[0.432, 0.568]^+$
Yi-1.5-6B	$+0.920\,[0.880, 0.960]^+$	$+0.956\,[0.920, 0.988]^+$	$+0.960\,[0.924, 0.992]^+$
Qwen2.5-7B	$+0.036\,[-0.016, 0.088]$	$+0.068\,[0.008, 0.128]^+$	$+0.016\,[-0.040, 0.072]$
pooled	$+0.765\,[0.748, 0.782]^+$	$+0.776\,[0.760, 0.792]^+$	$+0.775\,[0.758, 0.791]^+$

D_u, D_v and the remaining contrasts show no resolved change — a change interval that includes zero is no resolved change, never equivalence. The parser and the judge disagree on 1.2% of the $138{,}000$ responses, always where the parser counts an answer that the judge reads as an abstention or an invalid output; within-run consistency on the freshly judged pairs is $253/300 = 0.843\ [0.798, 0.880]$. The judged-subset comparison (Table A25) retains the three core arms; its earlier dose-arm rows are superseded by these full-panel values and omitted; the additional seeds (17/23) and OLMo-2-7B remain parser-scored robustness analyses.

Answer matching. Correctness uses two fixed answer matchers: a *relaxed* primary matcher, which accepts an output that equals a gold answer after normalization (lower-casing, punctuation and article removal, whitespace collapsing), or contains the normalized gold in a prediction at most eight tokens longer than the gold, or reaches token-level F1 ≥ 0.6 with it, or, for yes/no golds, begins with the gold token; and an alternative *containment* matcher, which accepts exactly when the normalized

Table A24: Parser-versus-judge disagreement on the non-empty responses, $d/(n-e)$ with a Wilson 95% interval on that denominator, pooled over a reader's strata of each kind: **(a)** adapter on — each reader's six Recipe-A insufficient-stratum cells (three seeds $\times$ two renderings; 27 series in total across the fleet) and the Recipe-B insufficient cells; **(b)** adapter off, by evidence state. Empty responses are counted separately (supplementary per-series table); a kind with no non-empty response is not assessed. † The SmolLM2 row in panel (b) reports a separate native-template baseline check; all other rows use the original fleet readouts. The fleet DiDs and aggregate disagreement rates retain the original readouts.

(a) adapter on, insufficient stratum

Reader	Recipe A	Recipe B
Qwen2.5-0.5B	70/1500 = 0.047 [0.037, 0.059]	61/1500 = 0.041 [0.032, 0.052]
Llama-3.2-1B	57/1500 = 0.038 [0.029, 0.049]	22/1500 = 0.015 [0.010, 0.022]
Qwen2.5-1.5B	22/1500 = 0.015 [0.010, 0.022]	316/1500 = 0.211 [0.191, 0.232]
SmolLM2-1.7B	45/1500 = 0.030 [0.022, 0.040]	64/1500 = 0.043 [0.034, 0.054]
Gemma-2-2B	26/1500 = 0.017 [0.012, 0.025]	10/1500 = 0.007 [0.004, 0.012]
Llama-3.2-3B	40/1500 = 0.027 [0.020, 0.036]	81/1500 = 0.054 [0.044, 0.067]
Phi-3.5-mini	86/1494 = 0.058 [0.047, 0.071]	407/1455 = 0.280 [0.257, 0.303]
Yi-1.5-6B	23/1500 = 0.015 [0.010, 0.023]	100/1500 = 0.067 [0.055, 0.080]
Qwen2.5-7B	91/1500 = 0.061 [0.050, 0.074]	123/1500 = 0.082 [0.069, 0.097]

(b) adapter off

Reader	insufficient stratum	sufficient stratum
Qwen2.5-0.5B	108/500 = 0.216 [0.182, 0.254]	54/500 = 0.108 [0.084, 0.138]
Llama-3.2-1B	116/500 = 0.232 [0.197, 0.271]	104/500 = 0.208 [0.175, 0.246]
Qwen2.5-1.5B	101/500 = 0.202 [0.169, 0.239]	30/500 = 0.060 [0.042, 0.084]
SmolLM2-1.7B†	143/500 = 0.286 [0.248, 0.327]	91/500 = 0.182 [0.151, 0.218]
Gemma-2-2B	176/498 = 0.353 [0.313, 0.396]	60/500 = 0.120 [0.094, 0.151]
Llama-3.2-3B	189/500 = 0.378 [0.337, 0.421]	57/500 = 0.114 [0.089, 0.145]
Phi-3.5-mini	140/500 = 0.280 [0.242, 0.321]	18/500 = 0.036 [0.023, 0.056]
Yi-1.5-6B	71/500 = 0.142 [0.114, 0.175]	21/500 = 0.042 [0.028, 0.063]
Qwen2.5-7B	359/500 = 0.718 [0.677, 0.756]	105/500 = 0.210 [0.177, 0.248]

Table A25: Core arms (A0, MARG-CE, CROSS-CE) on the judged subset of the primary 750-question panel (seed 7; 100 questions per arm, all four cells): sufficiency-oriented accuracy and the additive-probit coefficients under the frozen parser (L1) and the semantic reading (L3) on the same questions. Under L3 a sufficient-cell success is a committed, reference-correct answer and an insufficient-cell success an explicit abstention (an invalid output is neither). A fit is shown only when the scoring converged with no saturated cell and every coefficient below 5 in magnitude (*w.*: withheld). Subset comparisons, not full-panel values; the dose arms' judged-subset check is superseded by the full-panel re-scoring of Table A26 and omitted here.

Reader	Arm	n	SOA (L1)	SOA (L3)	u / v (L1)	u / v (L3)
Llama-3.2-3B	A0	100	0.395	0.403	*w.*	*w.*
Llama-3.2-3B	MARG-CE	100	0.760	0.760	$+1.18$ / $+0.03$	$+1.19$ / $+0.01$
Llama-3.2-3B	CROSS-CE	100	0.765	0.765	$+1.24$ / $+0.07$	$+1.24$ / $+0.07$
Qwen3-8B	A0	100	0.470	0.480	*w.*	*w.*
Qwen3-8B	MARG-CE	100	0.833	0.838	$+1.41$ / -0.04	$+1.47$ / -0.05
Qwen3-8B	CROSS-CE	100	0.838	0.830	*w.*	*w.*

gold answer appears in the output, with no length allowance. Neither acceptance set contains the other: containment accepts a long gold-containing output that every relaxed branch rejects, and the relaxed matcher accepts a non-containing output through token-level F1 — the eight-token allowance bounds only the relaxed matcher's own containment branch. To check the relaxed matcher we validate its decisions against two independent LLM judges (the *reference judges*); agreement is high — Cohen's $\kappa = 0.9378$, class-balanced accuracy 0.9314 $[0.8743, 0.9741]$, false positive rate 0.0704 $[0.0305, 0.1120]$, and false negative rate 0.0649 $[0, 0.1797]$.

Environment. The construction, training, readout and analysis code, with pinned Python/PyTorch environment specifications, is available in the anonymized repository (see also the reproducibility statement); the package versions recorded by the training runs are listed in Table A20.

Table A26: The mechanism coefficients under the frozen parser (L1) and the full semantic re-scoring (L3) of the same seed-7 dose, decision-loss and give-back readouts (audited 750-question panel, all four cells; the body's estimator with only the answer label changed; paired question-cluster bootstrap 95% intervals). **(a)** Dose ladder: v_{policy}-versus-LLR slope and Δu (MARG$\rightarrow$ g=0.95); the Spearman trends and crossover arms are identical under both readings. **(b)** Decision-loss control (Llama-3.2-3B): the L_{dec}-minus-L_{rest} endpoint contrasts. **(c)** Give-back cells (steps 48$\rightarrow$249; Llama = Llama-3.2-3B, Qwen = Qwen3-8B): every cell keeps $\Delta U > 0$ and $\Delta V < 0$ under the semantic reading.

(a) Dose ladder, seed 7

Reader	slope (L1)	slope (L3)	Δu (L1)	Δu (L3)
Llama-3.2-3B	$+0.410$ $[+0.386, +0.443]$	$+0.382$ $[+0.360, +0.410]$	-0.412 $[-0.506, -0.292]$	-0.458 $[-0.541, -0.362]$
Qwen3-8B	$+0.330$ $[+0.306, +0.363]$	$+0.312$ $[+0.290, +0.345]$	-0.393 $[-0.484, -0.279]$	-0.433 $[-0.519, -0.326]$

(b) Decision-loss control

Seed	D_u (L1)	D_u (L3)	D_v (L1)	D_v (L3)
7	$+0.522$ $[+0.424, +0.637]$	$+0.479$ $[+0.392, +0.577]$	$+0.013$ $[-0.089, +0.130]$	-0.008 $[-0.100, +0.092]$
17	$+0.494$ $[+0.389, +0.607]$	$+0.462$ $[+0.360, +0.570]$	-0.255 $[-0.361, -0.145]$	-0.278 $[-0.381, -0.173]$

(c) Give-back cells, steps 48$\rightarrow$249

Reader / g / seed	ΔU (L1)	ΔU (L3)	ΔV (L1)	ΔV (L3)
Llama / 0.5 / 7	$+0.725$ $[0.654, 0.800]$	$+0.701$ $[0.629, 0.776]$	-0.485 $[-0.555, -0.415]$	-0.479 $[-0.549, -0.411]$
Llama / 0.5 / 17	$+0.661$ $[0.594, 0.736]$	$+0.671$ $[0.606, 0.746]$	-0.201 $[-0.266, -0.130]$	-0.148 $[-0.210, -0.080]$
Llama / 0.8 / 7	$+0.499$ $[0.355, 0.604]$	$+0.443$ $[0.301, 0.546]$	-0.856 $[-1.014, -0.749]$	-0.851 $[-1.009, -0.742]$
Llama / 0.8 / 17	$+0.532$ $[0.386, 0.648]$	$+0.503$ $[0.352, 0.621]$	-0.483 $[-0.632, -0.380]$	-0.505 $[-0.655, -0.400]$
Llama / 0.9 / 7	$+0.482$ $[0.408, 0.574]$	$+0.480$ $[0.407, 0.572]$	-0.547 $[-0.636, -0.445]$	-0.544 $[-0.631, -0.443]$
Llama / 0.9 / 17	$+0.349$ $[0.258, 0.440]$	$+0.392$ $[0.307, 0.482]$	-0.510 $[-0.610, -0.415]$	-0.317 $[-0.412, -0.226]$
Qwen / 0.8 / 7	$+0.391$ $[0.268, 0.471]$	$+0.389$ $[0.269, 0.468]$	-0.374 $[-0.497, -0.297]$	-0.357 $[-0.479, -0.280]$

M Datasets and Evaluation Splits

Training data derive from HotpotQA (Yang et al., 2018); the controlled evaluation panels (250- and 750-question, the latter audited from 754 constructed sets) derive from 2WikiMultiHopQA (Ho et al., 2020), with a disjoint replication panel built from the MuSiQue-Ans development split (Trivedi et al., 2022) (2,391 constructed, 2,382 after the audit) and held out as test data throughout this paper; the external checkpoints use the ContractNLI (Koreeda & Manning, 2021) and ShARC (Saeidi et al., 2018) splits of Table 1 (their 2Wiki readouts are in Table A5). These datasets are disjoint, and the external data are never used to train. The one place they are tuned on is the inference stage of Section 6.2, whose thresholds and single-format choice are fitted on ContractNLI *development* documents and read on held-out development documents; the test splits are used only for evaluation. Primary analyses use development splits. Test data are used in five analyses, all on frozen policies: the cross-dataset mitigation replication of Section 6.1 on MuSiQue; the unseen-renderer generalization of Section C on MuSiQue; the MuSiQue readout of the controlled corpus intervention (Section 5.2); held-out confirmation for ChatQA-1.5-8B on the ContractNLI and ShARC test splits, reported in Section D; and the frontier-system evaluation of Section F on the ContractNLI and ShARC test splits, whose decision rules were fixed before any generation. The corpus feature audit of Section 5.2 inspects the publicly released ChatQA and TrustAlign post-training resources as data; the controlled intervention of Section D additionally builds its two matched training corpora from the public TrustAlign recipe data and trains only the TrustAlign-recipe intervention pair on them, never on any evaluation panel.

N External-LLM Audit of the Unanswerable Panels

We audit every candidate-insufficient evidence set used in the experiments: all 754 2Wiki and 2,391 MuSiQue sets. The two primary judges are OpenAI GPT-5.6 and Anthropic Claude Opus 5; a second OpenAI GPT-5.6 configuration independently adjudicates disagreements (all three accessed through the providers' APIs in August 2026). These models are external to the trained fleet and core readers; the OpenAI configurations are external to every evaluated system, while the Anthropic judge shares

a label with a frontier system evaluated in Section 4.2 (the overlap and its mitigation are stated in Section F). Each judge sees only the question, accepted answers, and remaining evidence; dataset identity, removed support, and the intended policy label are hidden. A judgment that the set is answerable must cite the supporting text in the displayed evidence.

Decision rule and adjudication. Each judge labels a set *truly insufficient*, *residually sufficient* — the remaining evidence still determines an accepted answer, supported by a citation — or *ambiguous*. The two primary judges label independently and a set is resolved when they agree; every disagreement is decided by the third configuration (47 of the 3,345 judged units, 1.41%, most of them on the calibration controls below). Raw agreement between the primary judges is 99.68% over the 3,145 audited sets, with Cohen's $\kappa = 0.628$; the two statistics diverge because residual sufficiency is rare, so a handful of disagreements moves κ far more than raw agreement, and we report both. No audited set resolved as ambiguous, so the truly-insufficient analysis and the analysis that also admits ambiguous sets coincide.

Positive controls and sensitivity. To calibrate the jury we add 200 blinded controls drawn from the *sufficient* cells (100 per dataset), presented identically and expected to read as answerable. Sensitivity is 0.75 $[0.657, 0.825]$ on 2Wiki and 0.84 $[0.756, 0.899]$ on MuSiQue (Wilson 95%): the jury is conservative, recovering most genuinely sufficient evidence while missing some. The detected residual-sufficiency rate below is therefore a lower-bound-style measurement, and the conservatism biases the audit toward keeping doubtful sets in the trusted panel rather than over-cleaning it.

Census and exclusions. The audit identifies thirteen candidate-insufficient sets as answerable: four from 2Wiki (0.53%, Wilson $[0.21\%, 1.36\%]$) and nine from MuSiQue (0.38%, $[0.20\%, 0.71\%]$), so over 99% of each constructed panel is confirmed unanswerable. We exclude all thirteen from the behavioral analyses, leaving audited unanswerable panels of 750 2Wiki and 2,382 MuSiQue sets (and 250 questions in the nine-reader subset). The excluded question identifiers are, for 2Wiki, `268337a00bdb11eba7f7acde48001122`, `495e71a20bb011ebab90acde48001122` (the one inside the 251-question fleet subset), `577638aa0bde11eba7f7acde48001122` and `7032bda20baf11ebab90acde48001122`, and, for MuSiQue, `2hop__21103_16254`, `2hop__31091_31122`, `2hop__67660_67223`, `2hop__82858_654855`, `3hop1__324448_831637_91775`, `3hop1__756602_831637_91775`, `3hop1__765847_831637_91775`, `3hop1__786632_831637_91775` and `4hop1__88342_75218_128008_89859`. These audited panels are the evaluation panels used throughout the paper.

Robustness to undetected residual answerability. The jury's positive-control sensitivity (0.75 on 2Wiki and 0.84 on MuSiQue) leaves open the possibility of undetected residual answerability. We therefore stress-test the audited 2Wiki panels under worst-case contamination. Suppose that up to a fraction ϵ of qids in an audited constructed-insufficient 2Wiki panel is in fact answerable and, adversarially at every ϵ, take the qids most favorable to the gate, crediting their renderer-induced answers as legitimate rather than unsupported. With those qids' unsupported-answer contributions set to zero while retaining the audited-panel denominator, the paired qid-bootstrap 95% lower endpoint of the pooled insufficient-stratum gate remains far above zero throughout $\epsilon \in [0, 5\%]$. At the sensitivity-adjusted reference level (0.71%) the gates are essentially unchanged. Even at $\epsilon = 5\%$ — about ten times the detected 0.53% rate — the lower endpoints are 0.940 and 0.857 for the frozen-A0 chain-minus-raw answer-rate gates of the core arms (Llama-3.2-3B and Qwen3-8B, seed 7, parser reading), and for the nine-reader pooled adapter DiD of the deterministic Recipe-A re-runs they are 0.610, 0.654 and 0.617 at seeds 7, 17 and 23 under the semantic reading (0.693, 0.736 and 0.703 under the parser). Per reader, eight of the nine keep a positive lower endpoint at every seed under both readings (the smallest, Phi-3.5-mini's, is 0.31 under the semantic reading); Qwen2.5-7B — unresolved at two of the three seeds — has lower endpoints of -0.080, $+0.032$ and -0.096. This is a conditional qid-level sensitivity result for the audited 2Wiki panels and the listed estimands: the sensitivity-derived levels are reference points rather than upper bounds on true contamination, and contamination changes the unsupported-answer interpretation rather than the observed renderer-induced answer-rate shift. The same adversarial analysis, extended to the strict 2,382-question MuSiQue test panel (detected residual rate 0.38%, jury sensitivity 0.84, sensitivity-adjusted reference 0.45%), leaves the audited MuSiQue endpoints — the mitigation contrasts and the frozen-A0 gates —

intact: all twelve treatment-versus-A0 mitigation contrasts keep adversarial 95% lower bounds above zero throughout $\epsilon \in [0, 5\%]$ — at $\epsilon = 5\%$ the lower bounds span 0.165–0.231 against unstressed contrasts of 0.197–0.264 — under a contrast-specific worst case in which a reclassified question's insufficient cells are re-scored under both coherent answerable-item scorings (answered; answered and reference-correct), taking the more adverse, with both the replacement and the trim-renormalize bounds; and the frozen-A0 chain-minus-raw gates keep large lower endpoints (0.947 and 0.829 at $\epsilon = 5\%$). Accordingly, within this stress range, residual answerability does not materially alter the magnitude or conclusion of the 2Wiki insufficient-stratum gate or of the MuSiQue mitigation replication.

O Additional Related Work

Beyond the accounts discussed in Section 2, several lines bear on the learning dynamics of the rendering gate. Qiu et al. (2024) give the closest account of feature learning under spurious correlations (interpretive context for Section 5.3). Simplicity bias (Shah et al., 2020; Kalimeris et al., 2019) and gradient starvation (Pezeshki et al., 2021) describe how easily decodable features can come to dominate training; the implicit bias of gradient descent (Soudry et al., 2018) shapes which of several separating solutions is reached; and spurious features drive characteristic out-of-distribution failures (Nagarajan et al., 2021). On the robustness side, group-robust training (Sagawa et al., 2020a;b) and selective prediction under group structure (Jones et al., 2021) operate downstream of group/support construction — consistent with our finding that group-robust reweighting adds no material SOA gain on top of an already-balanced renderer while improving accuracy and worst-group success under skew-correlated training data (Section I). Exact crossing (Section 6.1) instantiates counterfactual balancing (Kaushik et al., 2020; Veitch et al., 2021) on the renderer axis. Qian et al. (2024) probe trustworthiness-related representations across pre-training checkpoints; we track answer-versus-abstain dynamics during post-training.