BETTER ANSWERS, WEAKER EVIDENCE CONTROL: EVIDENCE BYPASS IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) can improve answer accuracy while weakening responsiveness to supporting evidence. We investigate this separation in scientific-document question answering through matched document interventions that change the supported answer while preserving the question and unrelated context. We identify evidence bypass: a student retains an originally correct answer after the document changes to support a different one. On QASPER with a Qwen3-4B student, standard OPD raises original-document accuracy by 3.4 percentage points, lowers counterfactual accuracy by 5.3 points, and increases bypass by 15.7 points; the same directional pattern appears on PeerQA. Controlled perturbations localize the shift to answer-critical evidence, and bypass grows with the student's pre-distillation closed-book preference for the original answer. A factorized study of teacher targets, privileged information, divergence, and supervision density finds that explicitly supplying the teacher with annotated answer-critical spans yields the largest joint improvement in evidence response and claim support. These findings motivate Evidence-Paired OPD (EP-OPD), which combines on-policy supervision on matched documents supporting different answers to the same question with an answer-level margin that favors each document's supported answer over its paired alternative. Across both datasets and two student sizes, EP-OPD recovers 54–66% of standard OPD's counterfactual-accuracy loss while preserving 80–85% of its original-document accuracy gain. Our results establish evidence responsiveness as a distinct training and evaluation target alongside answer accuracy.

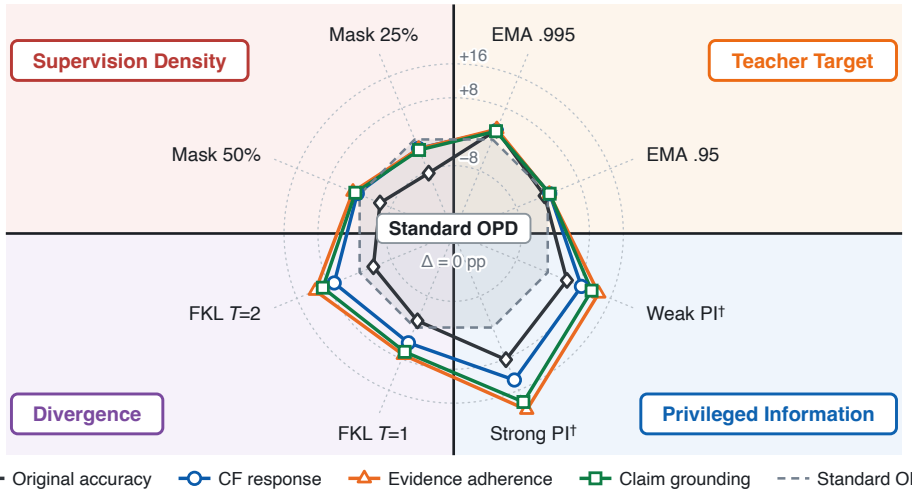

Figure 1: **A coordinate view of how ordinary OPD choices affect document use.** We vary supervision density, teacher target, divergence, and teacher privileged information one factor at a time around standard OPD (dashed contour). Each contour plots one metric's change (pp): original accuracy and the counterfactual response, evidence adherence, and claim grounding of the three panels of Figure 3; outward is better. Teacher and density changes move the three evidence metrics almost identically, so those contours overlap. PI (†): teacher privileged information.

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.

1 INTRODUCTION

Research on reasoning faithfulness distinguishes answer correctness and explanation plausibility from whether a model's stated reasons reflect the factors driving its decisions (Ye & Durrett, 2022; Turpin et al., 2023; Chen et al., 2025). Correct answers and coherent chains of thought (CoT) can coexist with a mismatch between stated justifications and actual decision influences. Interventions on reasoning traces and task inputs probe this relationship by measuring how predictions respond to changes in their stated basis (Lanham et al., 2023; Atanasova et al., 2023). Reasoning distillation brings this issue into training. Students learn from teacher-generated rationales and CoT traces (Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023), raising a further question: does transferring reasoning behavior also transfer dependence on supporting evidence?

On-policy distillation (OPD) provides a setting for studying this question: the student generates its own trajectories, and the teacher supervises the prefixes it visits (Agarwal et al., 2024; Gu et al., 2024). In document-grounded tasks, these trajectories bring the student's existing answer preferences into contact with evidence-conditioned teacher supervision. We investigate this question in scientific-document QA. Document evidence and parametric answer preferences often support the same answer on ordinary examples. We examine their relationship through controlled document variation: we change one answer-critical fact and update dependent mentions, producing a matched document that supports a different answer. Fresh responses to both versions measure how distillation changes the student's responsiveness to supporting evidence.

We identify evidence bypass: a model retains an originally correct answer after the document changes to support a different one. On QASPER with a Qwen3-4B student, standard OPD increases original-document accuracy from 49.0% to 52.4%, while counterfactual accuracy falls from 44.5% to 39.2% and bypass rises from 20.1% to 35.8%. The same directional pattern appears on PeerQA. These results reveal a separation between improvements in answer accuracy and changes in evidence dependence during distillation.

Further analysis connects this pattern to answer-critical information and pre-existing answer preferences. Paraphrases and irrelevant insertions preserve most of OPD's original-document gain, while evidence deletion reduces correct abstention. To examine how the training signal shapes evidence dependence, we vary four OPD design dimensions: teacher target, teacher privileged information, divergence loss, and supervision density. Among these interventions, providing the teacher with the annotated answer-critical span yields the largest gains in counterfactual accuracy and claim support, together with the largest reduction in bypass.

These findings motivate Evidence-Paired OPD (EP-OPD), which makes document-dependent answer variation explicit during training. Each question is paired with two documents supporting different answers. Both receive on-policy teacher supervision, and a paired answer-level objective encourages the student to favor each document's supported answer over its paired alternative. Across QASPER and PeerQA and two student sizes, EP-OPD recovers 54–66% of standard OPD's counterfactual-accuracy loss while preserving 80–85% of its original-document accuracy gain. These results establish evidence dependence as a training and evaluation target alongside answer accuracy.

Our contributions are threefold:

- We identify evidence bypass in document-grounded OPD through matched evidence interventions and characterize its relationship to the student's pre-distillation answer preferences.
- We organize OPD design along teacher target, teacher privileged information, divergence loss, and supervision density, showing how each shapes accuracy, evidence response, and claim support.
- We introduce Evidence-Paired OPD, which couples on-policy supervision with document-dependent answer preferences to strengthen evidence response while preserving most of distillation's accuracy gains.

2 RELATED WORK

2.1 CHAIN-OF-THOUGHT FAITHFULNESS AND EVIDENCE INTERVENTIONS

A correct answer accompanied by a plausible chain of thought (CoT) does not establish that the stated reasoning faithfully reflects the model's decision process. Ye & Durrett (2022) show that generated

explanations can lack input grounding. Turpin et al. (2023) demonstrate that CoT traces can omit biasing features influencing answers, while Chen et al. (2025) document analogous omissions of influential hints in reasoning models. Interventions on reasoning traces and task inputs probe whether predictions depend on their stated basis (Lanham et al., 2023; Atanasova et al., 2023), complementing evaluations of rationale sufficiency, comprehensiveness, and explanation–label association (DeYoung et al., 2020; Wiegreffe et al., 2021). Counterfactual data and contrast sets probe decision rules through targeted input changes (Kaushik et al., 2020; Gardner et al., 2020), while attribution research assesses whether claims are supported by source material (Rashkin et al., 2023; Gao et al., 2023). We apply these perspectives to scientific-document QA using answer-changing document pairs, evidence deletion, and claim-support evaluation. These diagnostics measure behavioral evidence responsiveness rather than directly verify internal reasoning.

2.2 On-policy and reasoning distillation

Reasoning distillation transfers teacher behavior through generated rationales and CoT traces (Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023). On-policy distillation (OPD) supervises student-generated trajectories: GKD applies teacher distributions at student-visited states (Agarwal et al., 2024), MiniLLM optimizes reverse KL using on-policy samples (Gu et al., 2024), and Li et al. (2026b) analyze progressive teacher–student alignment along these trajectories. CoT faithfulness raises a further question: does imitating teacher reasoning also transfer dependence on supporting evidence? Agreement on observed answers and traces alone need not establish such dependence. SCOTT addresses a related rationale-level issue through counterfactual rationale–answer pairs (Wang et al., 2023). We instead examine document-level evidence dependence under OPD, varying teacher target, teacher privileged information, divergence loss, and supervision density. Evidence-Paired OPD (EP-OPD) trains on matched documents supporting different answers to the same question, coupling answer preferences to source evidence rather than varying supplied rationales.

2.3 Context faithfulness and parametric knowledge

Knowledge-conflict research examines how models balance supplied context against parametric knowledge. Entity substitution and counterfactual QA probe reliance on these sources (Longpre et al., 2021; Neeman et al., 2023). FaithEval evaluates context-following under controlled evidence conditions (Ming et al., 2025), and Context-DPO trains preferences for context-faithful answers (Bi et al., 2025). Scientific-reasoning research further shows that stronger reasoning can make existing parametric knowledge more accessible (Li et al., 2026a). We connect these observations to distillation by relating the base student's closed-book answer preference to post-OPD responses under paired document interventions, without interpreting this association as direct identification of an internal reasoning mechanism.

3 Experimental Framework

We separate a model's accuracy on the supplied document from its response when answer-critical evidence changes, and compare how training alters each. This section defines the matched document interventions, the metrics and statistical comparisons built on them, and the models and OPD design variations we compare.

3.1 Datasets and Document Interventions

We evaluate on two scientific document QA datasets: QASPER comprises 5,049 questions over 1,585 NLP papers (Dasigi et al., 2021), and PeerQA comprises 579 questions drawn from reviews of 208 papers (Baumgaertner et al., 2025). Both provide answers and supporting evidence, with distinct question-collection procedures and answer formats. Splits are disjoint by source paper. The primary evaluation includes answerable questions whose answers can be represented as an entity, value, polarity, comparison, or bounded set; unanswerable examples contribute to abstention evaluation. We use full papers when they fit the context window. Otherwise, a fixed document pack combines annotated evidence with retrieval-ranked distractor paragraphs and is shared across methods.

We measure document response by changing the evidence that determines an answer while keeping the question fixed. For each eligible example, we edit one answer-critical span, changing an entity, value, comparison, polarity, or reported direction. We update dependent mentions consistently and preserve the question, section order, and unrelated context:

$$(q, D^{\mathrm{orig}}, a^{\mathrm{orig}}) \longrightarrow (q, D^{\mathrm{cf}}, a^{\mathrm{cf}}), \qquad a^{\mathrm{orig}} \neq a^{\mathrm{cf}}. \tag{1}$$

We generate a fresh response to each document version and validate each held-out pair so that the edited document uniquely supports its new answer. We also construct a meaning-preserving paraphrase D^{para}, an evidence-deleted version D^{drop}, and a length-matched irrelevant insertion D^{irr}. The answer-changing edit preserves answer type and scientific syntax and updates dependent mentions. Validators check label uniqueness, residual support for the original answer, and edit locality; held-out pairs also receive manual review before model evaluation. These versions measure response to semantic preservation, evidence availability, distraction, and answer-changing evidence (Section C).

3.2 Evidence-Response Metrics and Evaluation

Three primary quantities summarize answer behavior. *Original accuracy* measures correctness on D^{orig}; *counterfactual accuracy* (CFAcc) measures correctness on D^{cf}. *Bypass* is computed over the examples each model answers correctly on D^{orig}: the fraction on which the same model retains the original answer on D^{cf}. Because this conditioning set is model-specific, bypass differences compare conditional rates rather than changes on a fixed question set. The three measures capture performance on the observed document, adaptation to changed evidence, and persistence of the original answer. CFAcc and bypass are complementary: avoiding the original answer does not by itself mean producing the answer the edited document supports.

We also measure correct abstention after evidence deletion. We score original accuracy, counterfactual accuracy, and bypass with deterministic, item-specific answer canonicalizers. We split responses into independently evaluable claims and assess source support. A blinded evaluator receives the question, one randomized claim, its cited spans, and minimal neighboring context, with method and outcome metadata hidden. Labels are supported, contradicted, absent from the source, invalid composition, or irrelevant. Claim support precision (CSP) is the fraction of answer-relevant claims labeled supported. Experts evaluate a disjoint subset using the same units and definitions. Segmentation rules, prompts, and evaluation controls appear in Section D. Section B.1 gives the metric definitions, paired interaction estimand, and regression form.

Each matched question pair is the comparison unit, and source papers define resampling clusters. We aggregate claim labels by question and report scores averaged across seeds. The primary comparison estimates how the training-induced change in correctness differs between counterfactual and original documents using $10{,}000$ paired, paper-clustered bootstrap replicates. We apply Holm correction to secondary outcomes within each result family.

3.3 Models and Factorized OPD Design

The phenomenon analysis uses Qwen3-4B base students and a frozen Qwen3-30B-A3B-Instruct-2507 teacher (Yang et al., 2025). Training comparisons span the pretrained BASE checkpoint, answer-only supervision (ANS-SFT), imitation of teacher reasoning traces whose final answers pass the task verifier (answer-filtered CoT-SFT), and standard reverse-KL OPD. Principal runs share the question pool, optimizer-step count, retrieved context, and supervised output-token budget. We update all student parameters and run three seeds per principal configuration. Hyperparameters and compute matching appear in Section A.

We factorize standard OPD along four dimensions, teacher target, teacher privileged information, divergence, and supervision density, and vary one dimension at a time under an otherwise shared training setup. Teacher target: fixed reference or EMA with $\mu \in \{0.995, 0.95\}$. Teacher privileged information: gold evidence IDs or the answer-critical span. Divergence: reverse KL or forward KL with $T \in \{1, 2\}$. Supervised positions: all prefixes or the top 50% or 25% by pre-update student entropy. Questions, rollout settings, optimizer schedules, and student inputs remain fixed within each comparison. Exact objectives appear in Section A.4.

Table 1: QASPER with Qwen3-4B-Base (%). ↑/↓: higher/lower is better. Among the baseline training conditions, standard OPD attains the highest original and paraphrase accuracy while every evidence-sensitive outcome moves away from the base model. EP-OPD (last row; Section 6) partially recovers these outcomes at a 0.5 pp original-accuracy cost.

Method	Orig. ↑	Para. ↑	CFAcc ↑	Bypass ↓	Drop abst. ↑	CSP ↑
BASE	49.0	48.7	**44.5**	**20.1**	**22.0**	**69.8**
ANS-SFT	50.2	49.9	43.8	23.7	19.2	67.5
Answer-filtered CoT-SFT	51.4	51.0	42.3	29.6	16.0	64.2
Standard OPD	**52.4**	**51.9**	39.2	35.8	14.4	57.0
EP-OPD (ours)	51.9	51.6	42.7	25.9	19.1	65.4

Teacher privileged information is supplied only to the teacher; student rollouts and evaluation inputs never contain it. Entropy-mask controls hold rollouts and optimizer steps fixed, while the number of teacher-scored tokens intentionally varies with supervision density (Section A.4). Results of the factorized study are reported in Section 4.4 and Table 2.

4 EVIDENCE BYPASS UNDER ON-POLICY DISTILLATION

We examine how standard OPD changes evidence response in Qwen3-4B students, establishing the relationship between original-document accuracy and original-answer persistence under counterfactual evidence. We then characterize this change through document interventions and pre-distillation answer preferences, before examining how evidence response varies across the OPD design choices defined in Section 3.3.

4.1 ACCURACY GAINS ACCOMPANY STRONGER ORIGINAL-ANSWER PERSISTENCE

On QASPER, standard OPD increases original-document accuracy from 49.0% to 52.4%. On the matched answer-changing documents, counterfactual accuracy falls from 44.5% to 39.2%, and conditional bypass rises from 20.1% to 35.8%. The paired interaction is -8.7 percentage points (95% CI, $[-12.4, -5.1]$), capturing the separation between the training-induced changes on original and counterfactual documents. OPD therefore improves accuracy on the original documents while increasing persistence of the original answer when the supplied evidence supports a different one.

The supervised baselines place this pattern in the context of different training signals. Answer-only SFT produces modest increases in original accuracy and bypass; answer-filtered CoT-SFT produces larger increases in both; and standard OPD exhibits the largest original-document gain and bypass increase among these baseline training conditions (first four rows of Table 1). The same direction appears on PeerQA with the 4B student: standard OPD raises original accuracy by 2.1 percentage points and lowers counterfactual accuracy by 2.6 points, yielding a paired interaction of -4.7 points. Across both datasets, the matched evaluations reveal a consistent shift toward retaining the original answer after an answer-changing edit.

4.2 DOCUMENT INTERVENTIONS LOCALIZE THE SHIFT TO ANSWER-CRITICAL EVIDENCE

The document controls connect this shift to the information needed to answer the question. On QASPER, accuracy on meaning-preserving paraphrases rises from 48.7% to 51.9%, closely tracking the original-document improvement. Irrelevant insertions likewise preserve most of the original-document gain. Removing the necessary evidence produces a larger change: correct abstention falls from 22.0% to 14.4%. The accuracy gain thus carries over to edits that preserve the answer, whereas abstention under evidence deletion and accuracy under answer-changing edits both decline.

Claim-level evaluation extends the analysis to the support for individual answer-relevant statements. Automatic evidence error rises from 30.2% to 43.0%, and error on the blinded expert subset rises from 28.5% to 38.3%. These changes accompany a 12.8-point reduction in claim support precision. Table 7 brings together the document-operator and claim-level outcomes, linking the change in final-answer behavior to weaker support for the claims used in the response.

Table 2: Factorized OPD interventions on QASPER with Qwen3-4B (%). Each intervention changes one design dimension relative to standard OPD. †Gold evidence information is provided only to the teacher.

Axis	Single change	Orig. ↑	CFAcc ↑	Bypass ↓	CSP ↑
Reference	Standard OPD	52.4	39.2	35.8	57.0
Teacher	EMA, $\mu = 0.995$	54.8	41.5	33.2	59.1
Teacher	EMA, $\mu = 0.95$	51.6	39.7	35.4	57.5
Priv. info.	Weak PI: evidence IDs†	57.3	47.8	22.8	68.2
Priv. info.	Strong PI: critical span†	60.6	52.6	14.8	76.0
Divergence	Forward KL, $T = 1$	50.7	43.1	28.8	63.2
Divergence	Forward KL, $T = 2$	48.9	45.7	24.5	66.4
Density	Top-entropy 50%	47.2	39.8	34.1	58.1
Density	Top-entropy 25%	43.8	36.9	38.2	54.3

4.3 BYPASS INCREASES WITH PRE-DISTILLATION ANSWER PREFERENCE

The additional bypass induced by OPD grows monotonically across the resulting prior-strength quartiles, from approximately 5 percentage points in the weakest-prior quartile to 20 points in the strongest. Across the two datasets, the three-way interaction between OPD, counterfactual evaluation, and high prior strength lies between -7 and -10 points. The strongest change in evidence response thus occurs where the base student already favors the original answer.

The document interventions and prior stratification provide complementary views of evidence bypass: the former locate the shift in answer-critical evidence, and the latter relate its magnitude to an existing answer preference. We next examine how evidence response changes when individual OPD design choices are varied.

4.4 EVIDENCE RESPONSE UNDER OPD DESIGN VARIATIONS

The preceding analysis concerns standard OPD. We now ask whether the separation between original accuracy and evidence response persists, and how it changes, when a single training choice is varied. Using the factorized design in Section 3.3, we vary one dimension at a time on QASPER with Qwen3-4B. Table 2 reports the results, and Figure 1 gives an overview.

Teacher-target changes remain close to standard OPD: a slow EMA target reaches 54.8% original accuracy and 41.5% CFAcc, with 33.2% bypass and 59.1% CSP. A faster EMA target produces a smaller shift. Forward KL traces a different accuracy-response profile: moving from $T = 1$ to $T = 2$ raises CFAcc from 43.1% to 45.7% and reduces bypass from 28.8% to 24.5%, as original accuracy moves from 50.7% to 48.9%. Concentrating supervision on fewer high-entropy prefixes progressively lowers original accuracy.

Teacher privileged information produces the largest joint movement across the evidence-sensitive measures. Providing gold evidence IDs raises CFAcc to 47.8%, lowers bypass to 22.8%, and raises CSP to 68.2%. Providing the exact answer-critical span moves these quantities to 52.6%, 14.8%, and 76.0%, with original accuracy reaching 60.6%. Among the tested interventions, supplying answer-critical evidence to the teacher yields the largest joint improvement in CFAcc, bypass, and CSP.

5 EVIDENCE-PAIRED OPD

When each question is paired with only one answer-supporting document during training, evidence-responsive and document-insensitive strategies may behave identically on the observed examples (Proposition 1). We therefore pair each question with documents supporting different answers, so that document-insensitive answer preferences cannot simultaneously satisfy both supervision targets.

Evidence-Paired OPD (EP-OPD) turns the matched-document intervention into a training signal. Each question appears with two validated documents that support distinct answers. The student receives on-policy teacher supervision under each document, together with an answer-level objective that couples its answer preferences to the evidence supplied in that document. Figure 2 shows the document intervention and the associated training outcomes.

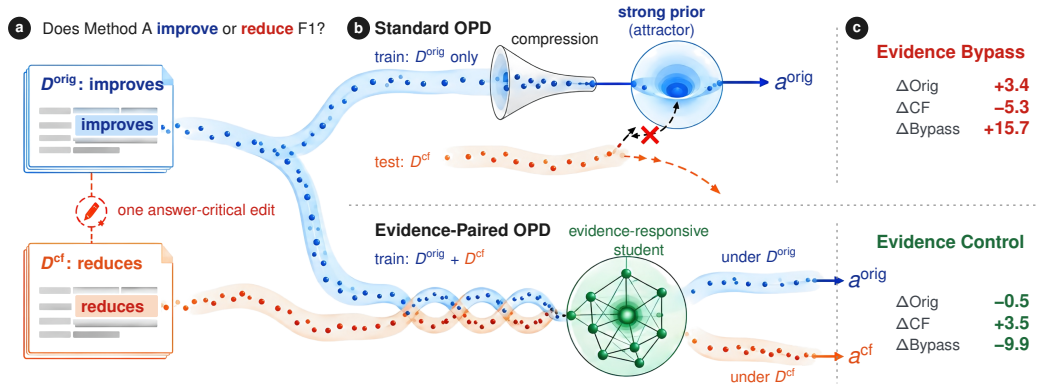

Figure 2: **Minimal evidence interventions identify bypass; paired evidence repairs it.** One answer-critical edit flips the supported answer while the question and document structure stay fixed. Standard OPD sees only D^{orig} and may preserve a^{orig} under D^{cf}; EP-OPD co-distills the pair and follows the edit. Values are Qwen–4B; deltas are relative to Base(top) and standard OPD(bottom).

5.1 Paired-document objective

For a validated pair, let $x^0 = (q, D^0)$ and $x^1 = (q, D^1)$ support answers $a^0 \neq a^1$. The student independently samples a response under each input. At every student-visited prefix, the frozen teacher conditions on the same question and document and supplies the next-token distribution used by the standard reverse-KL OPD objective.

Let $s_\theta(a \mid x)$ denote the length-normalized log probability of canonical answer a. We encourage each document to favor its supported answer over the answer associated with its partner:

$$\mathcal{L}_{\text{pair}} = \sum_{b \in \{0,1\}} \left[m - s_\theta(a^b \mid x^b) + s_\theta(a^{1-b} \mid x^b) \right]_+, \tag{2}$$

where $m > 0$ is the answer margin and $[z]_+ = \max\{z, 0\}$. The full objective is

$$\mathcal{L}_{\text{EP-OPD}} = \frac{1}{2} \sum_{b \in \{0,1\}} \mathcal{L}_{\text{OPD}}(x^b; \theta) + \lambda \mathcal{L}_{\text{pair}}.$$

The OPD term transfers teacher behavior along student-generated trajectories, and the paired term couples answer preferences to document variation. When the pair loss is zero, each supported answer exceeds its paired alternative in score by at least m under its own document. The two canonical answers consequently exchange their ranking across the pair. Appendix B.3 formalizes this property, and Section A.2 gives the token-level loss and answer-score definitions.

Algorithm 1 Evidence-Paired On-Policy Distillation

Require: Paired data $\mathcal{P}$, student π_θ, frozen teacher π_T.
Require: Margin m, weight λ, batch size B, learning rate η.
for each training iteration **do**
  Sample $\mathcal{B} \subset \mathcal{P}$, $|\mathcal{B}| = B$; set $\widehat{\mathcal{L}} \leftarrow 0$.
  for each $(q, D^0, a^0, D^1, a^1) \in \mathcal{B}$ **do**
    for $b \in \{0, 1\}$ **do**
      $x^b \leftarrow (q, D^b)$; sample $y^b \sim \pi_\theta(\cdot \mid x^b)$.
      $p_t \leftarrow \pi_\theta(\cdot \mid x^b, y^b_{<t})$, $r_t \leftarrow \pi_T(\cdot \mid x^b, y^b_{<t})$ for all t.
      $\ell_b \leftarrow \frac{1}{|y^b|} \sum_{t=1}^{|y^b|} \sum_{v \in \mathcal{V}} p_t(v) \log \frac{p_t(v)}{r_t(v)}$.
      for $c \in \{0, 1\}$ **do**
        $s_{bc} \leftarrow \frac{1}{|a^c|} \sum_{j=1}^{|a^c|} \log \pi_\theta(a^c_j \mid x^b, a^c_{<j})$.
    $\mathcal{L}_{\text{pair}} \leftarrow [m - s_{00} + s_{01}]_+ + [m - s_{11} + s_{10}]_+$.
    $\widehat{\mathcal{L}} \leftarrow \widehat{\mathcal{L}} + \frac{1}{B} \left[\frac{\ell_0 + \ell_1}{2} + \lambda \mathcal{L}_{\text{pair}} \right]$.
  Update θ using $\nabla_\theta \widehat{\mathcal{L}}$ at learning rate η.
Ensure: Updated student parameters θ.

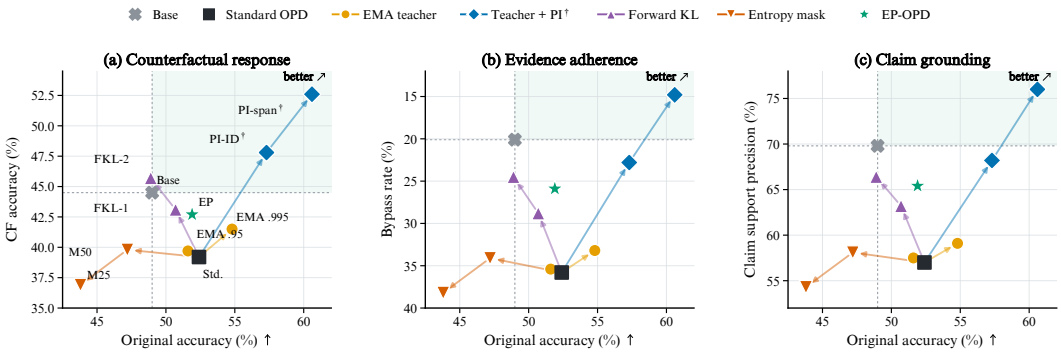

Figure 3: **One-axis OPD interventions trace different accuracy–evidence frontiers.** Panels pair original accuracy with CFAcc, reversed bypass, or CSP (top-right is better). Arrows change one factor from standard OPD; dashed guides mark Base. PI (†) is an oracle and EP-OPD a symmetric-context reference.

5.2 Training procedure

We construct and validate document pairs before optimization and place both members of each pair in the same mini-batch. Each member receives an independent student rollout and the corresponding teacher supervision. The paired answer scores are computed by teacher forcing over the short canonical answer strings. The margin and pair weight are selected on the QASPER development set and transferred unchanged to the second student size and PeerQA.

At inference, the trained student receives a single question–document input and generates its response. Algorithm 1 summarizes the optimization procedure. Appendix A specifies the hyperparameters and supervision-budget matching.

6 Evaluating Evidence-Paired OPD

We evaluate EP-OPD on held-out document pairs using the protocol in Section 3. The main comparison uses the Qwen3-4B student from the phenomenon analysis; the cross-dataset and scale study additionally includes Qwen3-1.7B. Details appear in Appendix A.3.

6.1 Answer accuracy and evidence response

On QASPER with Qwen3-4B (Table 1), EP-OPD achieves 51.9% original-document accuracy, preserving 85% of the gain obtained by standard OPD over the base student. The original-accuracy change relative to standard OPD is -0.5 percentage points, and paraphrase accuracy is 51.6%. Counterfactual accuracy increases by 3.5 points to 42.7%, recovering 66% of standard OPD's counterfactual-accuracy loss. Conditional bypass falls from 35.8% to 25.9%, a reduction of 9.9 points.

The improvement extends to evidence availability and claim support. Correct abstention after evidence deletion rises by 4.7 points to 19.1%, and claim support precision rises by 8.4 points to 65.4%. Paired training therefore improves adaptation to answer-changing evidence, abstention when necessary evidence is missing, and support for answer-relevant claims while preserving most of the original-document accuracy gain.

Figure 3 places this response profile alongside the factorized OPD controls from Section 4.4. Teacher targets, divergence choices, and supervision density yield different combinations of original accuracy and evidence response. The teacher-side privileged-information controls measure the effect of supplying annotated evidence directly to the teacher. EP-OPD introduces variation through matched training documents, with teacher and student sharing the document input for each member. Across the three panels, paired training improves counterfactual response, evidence adherence, and claim grounding relative to standard OPD.

Table 3: Cross-dataset and scale replication. Entries are percentage-point changes of standard OPD relative to the frozen base, except the final column, which is the fraction of OPD's CFAcc loss recovered by EP-OPD.

Dataset	Student	ΔOrig.	ΔCFAcc	CF recovery
QASPER	1.7B	+2.6	−3.4	59%
QASPER	4B	+3.4	−5.3	66%
PEERQA	1.7B	+2.9	−4.5	62%
PEERQA	4B	+2.1	−2.6	54%

6.2 CONSISTENCY ACROSS DATASETS AND STUDENT SIZES

Table 3 reports the standard-OPD change from the base student in each dataset–student setting and the fraction of its counterfactual-accuracy loss recovered by EP-OPD. On QASPER, paired training recovers 59% of the counterfactual loss for the 1.7B student and 66% for the 4B student. The corresponding recovery on PeerQA is 62% and 54%. Across the four settings, EP-OPD retains 80–85% of the original-document accuracy gain while recovering 54–66% of the counterfactual loss. The improvement in document response thus extends across both scientific QA datasets and both student sizes under the shared paired-training configuration.

6.3 EVIDENCE RESPONSE ACROSS PRIOR-STRENGTH STRATA

The frozen-base prior stratification connects the effect of pairing to the example-level pattern established in the phenomenon analysis. As shown in Appendix F.2, EP-OPD lowers bypass in every quartile. In the weakest-prior quartile, bypass falls from 17% under standard OPD to 14%; in the strongest, it falls from 49% to 35%. Relative to the base student, these values correspond to additional bypass of 2 and 6 percentage points, compared with 5 and 20 points under standard OPD. Pairing thus improves evidence following throughout the prior distribution, with the largest absolute reduction in the strongest-prior quartile.

Component ablation. We compare three paired-document variants on QASPER with Qwen3-4B: Augmented OPD, Paired ANS-SFT, and Augmented OPD with answer cross-entropy. More details in Appendix E. Relative to Augmented OPD, EP-OPD improves CFAcc by 4.2 percentage points and CSP by 3.1 points, with a 0.2-point decrease in original accuracy. Replacing answer cross-entropy with the paired margin improves CFAcc and CSP by 1.8 and 5.7 points, respectively, with original accuracy changing from 52.5% to 51.9%. Paired ANS-SFT attains the lowest bypass rate (25.3%), while EP-OPD achieves higher original accuracy, CFAcc, and CSP.

7 CONCLUSION

We identify evidence bypass in document-grounded on-policy distillation: students improve their accuracy on original documents while increasingly retaining the original answer when the document supports a different one. Matched interventions on QASPER and PeerQA reveal that this shift tracks answer-critical evidence and grows with the student's pre-distillation preference for the original answer. We decompose standard OPD along four dimensions—teacher target, teacher privileged information, divergence loss, and supervision density—and find that teacher-side oracle access to answer-critical evidence yields the largest joint improvement in evidence-sensitive outcomes.

Building on these findings, Evidence-Paired OPD trains on matched documents that support different answers to the same question, explicitly coupling answer preferences to document evidence. It improves counterfactual accuracy and claim support, reduces bypass, and preserves most of standard OPD's original-document accuracy gain. Collectively, these results establish evidence response as a distinct dimension of distillation quality and motivate training objectives that strengthen both answer accuracy and responsiveness to answer-relevant evidence.

AI USE STATEMENT

Generative AI tools were used solely for limited assistance with language polishing, formatting, routine coding, and debugging. The research, analysis, interpretation of results, and preparation of the final content were carried out by the human authors. All results and citations were independently verified by the authors, who take full responsibility for the work.

ETHICS STATEMENT

Counterfactual documents are controlled research stimuli, not claims about the cited papers. They are clearly marked in released metadata and must not be indexed or redistributed as authentic scientific text. Human validation uses only document excerpts and model outputs; no sensitive personal data are collected. The method may improve evidence following, but it does not guarantee factuality when supplied evidence is itself false or adversarial, so counterfactual compliance should not be treated as a universal trustworthiness criterion.

REPRODUCIBILITY STATEMENT

The paper specifies the document interventions, metrics, statistical unit, resampling protocol, datasets, models, and baselines (Section 3), the training objective and algorithm (Section 5), and the formal estimand (Section B.1). The appendix provides proofs, construction validators, evaluation prompts, hyperparameters, and additional result tables. The source package includes a summary result manifest and uses shared macros for the main reported outcomes. An anonymous implementation and generated intervention manifests will accompany the submission.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.25. URL https://aclanthology.org/2023.acl-short.25/.

Tim Baumgaertner, Ted Briscoe, and Iryna Gurevych. PeerQA: A scientific question answering dataset from peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 508–544. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.22.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.536.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. URL https://arxiv.org/abs/2505.05410.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.830.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507.

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. URL https://openreview.net/forum?id=Sklgs0NFvr.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Alan Li, Yixin Liu, Arpan Sarkar, Doug Downey, and Arman Cohan. Demystifying scientific problem-solving in LLMs by probing knowledge and reasoning. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2508.19202.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.151.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"? In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QYo5WZGI8k.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.559.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5546–5558. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.304.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.804. URL https://aclanthology.org/2021.emnlp-main.804/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30378–30392, 2022.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.307.

Table 4: Training and decoding configuration. Pair hyperparameters are selected once on QASPER development data and transferred unchanged.

Setting	Value
Student checkpoints	Qwen3-1.7B-Base; Qwen3-4B-Base
Frozen teacher	Qwen3-30B-A3B-Instruct-2507
Trainable parameters	all student parameters
Optimizer	AdamW, $\beta = (0.9, 0.95)$, weight decay 0.1
Peak learning rate	1.5×10^{-5} (1.7B); 1.0×10^{-5} (4B)
Schedule	3% warmup, cosine decay, two epochs
Global batch	128 documents (64 synchronized pairs for EP-OPD)
Maximum input / output	32,768 / 768 tokens
Training rollout	temperature 0.8, top-$p = 0.95$
Primary evaluation	greedy decoding
Gradient clipping / precision	1.0 / bfloat16
OPD divergence	token-average reverse KL, full vocabulary
Pair margin m / weight λ	0.20 / 0.50
Random seeds	3 per principal configuration

A TRAINING AND IMPLEMENTATION DETAILS

A.1 OPTIMIZATION

All methods use the same document serialization and answer delimiter. The frozen evidence-preserving pack keeps the retrieval channel fixed across methods, separating evidence use from retrieval-augmented generation (Lewis et al., 2020; Asai et al., 2024). OPD and EP-OPD compute the teacher distribution at every non-padding student token; no reward model or answer-only policy gradient is used. To match supervision tokens, we subsample longer off-policy traces and resample shorter answer-only targets. Training examples are shuffled at the paper level. Development selection uses original accuracy subject to a counterfactual-response constraint fixed before test evaluation; the test checkpoint is selected once per seed.

For EP-OPD, a batch contains the original and counterfactual member of each pair. The global batch size counts documents, so the number of pairs per update is half the number of standard-OPD documents. Both members receive independent on-policy rollouts. The margin score is teacher-forced only over the short canonical answer string and is length normalized as in Equation (2).

A.2 EVIDENCE-PAIRED OPD DETAILS

Let $x = (q, D)$ and let the student sample $y \sim \pi_\theta(\cdot \mid x)$. At each student-visited prefix $y_{<t}$, the frozen teacher provides a next-token distribution $\pi_T(\cdot \mid x, y_{<t})$. The standard baseline uses length-normalized reverse KL,

$$\mathcal{L}_{\mathrm{OPD}}(x;\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<t}) \,\|\, \pi_T(\cdot \mid x, y_{<t}))\right]. \tag{3}$$

For a validated pair $x^0 = (q, D^0)$ and $x^1 = (q, D^1)$, the answer score used in the margin is

$$s_\theta(a \mid x) = \frac{1}{|a|}\sum_{t=1}^{|a|} \log \pi_\theta(a_t \mid x, a_{<t}), \tag{4}$$

and the complete training objective is

$$\mathcal{L}_{\text{EP-OPD}} = \tfrac{1}{2}\sum_{b\in\{0,1\}} \mathcal{L}_{\mathrm{OPD}}(x^b;\theta) + \lambda\mathcal{L}_{\mathrm{pair}}, \tag{5}$$

with $\mathcal{L}_{\mathrm{pair}}$ defined in Equation (2).

The complete update is summarized in Algorithm 1 in the main text.

A.3 COMPUTE MATCHING

Teacher forward calls dominate OPD compute. Standard OPD therefore receives two independently sampled observational documents for each pair consumed by EP-OPD, equalizing teacher-scored tokens and optimizer steps. Reported method comparisons also match maximum input length, maximum generated length, decoding temperature, and checkpoint-selection schedule. Where EP-OPD is evaluated at both student sizes, the runs share data manifests and hyperparameters except for learning rate.

A.4 FACTORIZED OPD CONTROL OBJECTIVES

Every factor control starts from the same QASPER–4B training manifest and changes only the specified training choice. EMA targets are no-gradient moving teachers; if ϕ_k denotes the corresponding online teacher checkpoint after update k, their parameters obey

$$\bar{\phi}_k = \mu\bar{\phi}_{k-1} + (1-\mu)\phi_k, \qquad \mu \in \{0.995, 0.95\}, \tag{6}$$

with identical initialization and update ordering within the two EMA runs. The fixed-teacher row instead uses the external Qwen3-30B-A3B target throughout.

For privileged-context controls, the student's context remains $x = (q, D)$. The teacher receives $x_T = (q, D, z_E)$, where weak PI makes z_E the IDs of gold evidence paragraphs and strong PI makes z_E the exact answer-critical evidence span. PI is never present in the student rollout or at evaluation. Because these rows consume information unavailable to the symmetric-context methods, they are oracle interventions in the learning-with-privileged-information sense (Lopez-Paz et al., 2016) and estimate the value of resolving the teacher's evidence-location uncertainty.

At a student-visited prefix $s_t = (x, y_{<t})$, forward KL uses

$$\mathcal{L}_{\mathrm{FKL}}(s_t; T) = T^2 D_{\mathrm{KL}}\Big(\pi_T^{(T)}(\cdot \mid s_t) \,\|\, \pi_\theta^{(T)}(\cdot \mid s_t)\Big), \tag{7}$$

where both distributions use temperature T and the conventional T^2 factor compensates for temperature-dependent gradient scaling. Entropy masks are computed from the frozen pre-update student distribution for the complete rollout. Let $m_t^{(q)}$ select prefixes in the top $q \in \{0.50, 0.25\}$ entropy quantile. The masked loss is

$$\mathcal{L}_{\mathrm{mask}}^{(q)} = \frac{\sum_t m_t^{(q)} \mathcal{L}_{\mathrm{OPD}}(s_t)}{\sum_t m_t^{(q)} + \epsilon}. \tag{8}$$

Normalizing over the selected prefixes keeps the loss an average over supervised positions. The ablation therefore changes which positions receive supervision, rather than reducing the loss simply because fewer positions are selected. Rollouts and optimizer steps are held fixed; teacher-scored tokens intentionally fall with q because supervision density is the manipulated variable.

B COMPARISON AND SUPPORTING RESULTS

B.1 METRIC DEFINITIONS AND PAIRED TRAINING COMPARISON

Let $C_i^v(\theta) = \mathbf{1}[\hat{a}_\theta(q_i, D_i^v) = a_i^v]$ denote answer correctness on document version $v \in \{\mathrm{orig}, \mathrm{cf}\}$. The primary answer-level metrics are

$$\mathrm{Acc}_{\mathrm{orig}}(\theta) = \mathbb{E}_i[C_i^{\mathrm{orig}}(\theta)], \tag{9}$$

$$\mathrm{CFAcc}(\theta) = \mathbb{E}_i[C_i^{\mathrm{cf}}(\theta)], \tag{10}$$

$$\mathrm{BypassRate}(\theta) = \Pr\Big[\hat{a}_\theta(q_i, D_i^{\mathrm{cf}}) = a_i^{\mathrm{orig}} \mid C_i^{\mathrm{orig}}(\theta) = 1\Big]. \tag{11}$$

Original accuracy measures performance on the observed document, CFAcc measures adaptation to the answer-changing document, and bypass measures persistence of the original answer among examples answered correctly before the intervention.

For a training method M, with base parameters θ_0 and trained parameters θ_M, define

$$\Delta_M^v = \mathbb{E}_i[C_i^v(\theta_M) - C_i^v(\theta_0)]\,, \tag{12}$$

$$\Psi_M = \Delta_M^{\mathrm{cf}} - \Delta_M^{\mathrm{orig}}. \tag{13}$$

The paired quantity Ψ_M measures how the training-induced change differs between counterfactual and original documents. We report it together with the corresponding change in bypass to summarize how training changes document response.

B.2 Single-document ambiguity

Proposition 1 (Observational non-identifiability). *Let $\mathcal{S} = \{(q_i, D_i, a_i)\}_{i=1}^n$ be a training support in which each q_i occurs with one answer-supporting document. If an evidence-responsive policy $f(q, D)$ is correct on $\mathcal{S}$, then there exists a question-only policy $g(q)$ with identical predictions and identical empirical answer or trace-imitation risk on $\mathcal{S}$. Consequently, observational supervision alone does not identify evidence response.*

Proof. Let f be any evidence-responsive policy that produces a conditional output distribution $p_f(\cdot \mid q_i, D_i)$ on each observed example. Because each observed q_i is paired with a single D_i, define a question-only policy g on the empirical support by

$$p_g(\cdot \mid q_i) = p_f(\cdot \mid q_i, D_i). \tag{14}$$

Outside the observed questions, define g arbitrarily. Then f and g induce identical distributions over answers on $\mathcal{S}$, hence identical empirical answer loss for every proper scoring rule. If their next-token distributions are matched at every prefix reached on $\mathcal{S}$, they also induce the same trace-imitation and on-policy token risk there. Nevertheless, g is invariant to replacing D_i with any D_i', whereas f need not be. The two policies therefore agree observationally but differ under document intervention. □

The proposition characterizes the observed training support: one document per question leaves evidence-responsive and question-conditioned solutions observationally indistinguishable on that support.

Proposition 2 (Paired shortcut exclusion). *Consider a balanced pair with fixed q, distinct answers $a^0 \neq a^1$, and documents D^0, D^1. Any stochastic question-only policy $p(a \mid q)$ has average negative log-likelihood at least $\log 2$ on the pair. Any deterministic question-only policy has pair accuracy at most $1/2$.*

Proof. Write $p_0 = p(a^0 \mid q)$ and $p_1 = p(a^1 \mid q)$. Since $a^0 \neq a^1$, $p_0 + p_1 \leq 1$. The balanced pair negative log-likelihood is

$$\ell = -\tfrac{1}{2}(\log p_0 + \log p_1) = -\log \sqrt{p_0 p_1}. \tag{15}$$

By AM–GM, $\sqrt{p_0 p_1} \leq (p_0 + p_1)/2 \leq 1/2$, so $\ell \geq \log 2$. A deterministic question-only policy returns the same answer for both documents, so it can match at most one of the two distinct labels. □

For a balanced K-way evidence set with K distinct supported answers, the same argument gives NLL at least $\log K$ and deterministic accuracy at most $1/K$.

B.3 What the paired margin guarantees locally

Corollary 3 (Local evidence-response certificate). *If $\mathcal{L}_{\mathrm{pair}} = 0$, then the paired response*

$$R_\theta = \big[s_\theta(a^0 \mid x^0) - s_\theta(a^1 \mid x^0)\big] + \big[s_\theta(a^1 \mid x^1) - s_\theta(a^0 \mid x^1)\big] \tag{16}$$

satisfies $R_\theta \geq 2m$. Every question-only policy has $R_\theta = 0$, so it cannot attain zero pair loss for $m > 0$.

Proof. Zero hinge loss in Equation (2) implies

$$s_\theta(a^0 \mid x^0) - s_\theta(a^1 \mid x^0) \geq m, \tag{17}$$

$$s_\theta(a^1 \mid x^1) - s_\theta(a^0 \mid x^1) \geq m. \tag{18}$$

Summing yields $R_\theta \geq 2m$. For a question-only policy, $s_\theta(a \mid q, D^0) = s_\theta(a \mid q, D^1)$ for every a, making the two brackets exact negatives and $R_\theta = 0$. □

Corollary 4 (Paired answer-decision guarantee)**.** *Let $d(x) = \arg\max_{a \in \{a^0, a^1\}} s_\theta(a \mid x)$. If $\mathcal{L}_{\text{pair}} = 0$ and $m > 0$, then $d(x^0) = a^0$ and $d(x^1) = a^1$. Consequently, candidate-set accuracy is one on both documents and pairwise bypass is zero.*

Proof. Zero pair loss with $m > 0$ implies $s_\theta(a^0 \mid x^0) > s_\theta(a^1 \mid x^0)$ and $s_\theta(a^1 \mid x^1) > s_\theta(a^0 \mid x^1)$. The candidate-set argmax therefore selects the document-supported answer on both versions. □

These statements characterize the two canonical alternatives used by the margin. The main evaluation extends the analysis to fresh document pairs under free generation.

B.4 A SIMPLE ILLUSTRATION OF PRIOR CANCELLATION

Proposition 5 (Prior cancellation in a two-source model)**.** *Let $c \in \{-1, +1\}$ index which of two answers a document supports, and let the logit for the positive answer be $u(q) + cv$, where $u(q)$ is constant across the pair and $v > 0$ is the evidence contribution. For fixed v, the balanced paired logistic risk has the unique minimizer $u^* = 0$. Observational risk on only $c = +1$ depends on $u + v$ and therefore cannot identify their allocation.*

Proof. For the fixed question, write $u = u(q)$. The balanced paired logistic risk is

$$L(u, v) = \tfrac{1}{2} \log(1 + e^{-(u+v)}) + \tfrac{1}{2} \log(1 + e^{u-v}). \tag{19}$$

For fixed v, L is strictly convex in u, symmetric under $u \mapsto -u$, and has $\partial L / \partial u = 0$ at $u = 0$; hence its unique minimizer is $u^* = 0$. With only the positive document, the risk is $\log(1 + e^{-(u+v)})$, which is constant on every level set of $u + v$ and cannot identify how the logit is allocated. □

Writing $\sigma(z) = (1 + e^{-z})^{-1}$, original-answer persistence under the counterfactual document is $B(u, v) = \sigma(u - v)$. At fixed v,

$$\frac{\partial B}{\partial u} = \sigma(u - v)\big[1 - \sigma(u - v)\big] > 0, \tag{20}$$

$$\frac{\partial L}{\partial u} = \tfrac{1}{2}\big[\sigma(u - v) - \sigma(-u - v)\big] > 0 \qquad (u > 0), \tag{21}$$

$$\frac{\partial^2 L}{\partial u^2} = \tfrac{1}{2}\big[\sigma'(u - v) + \sigma'(-u - v)\big] > 0. \tag{22}$$

Thus a larger positive prior increases persistence and also strengthens the paired gradient opposing it in this toy model. This relation motivates the pre-training prior stratification used in the main experiments.

C EVIDENCE INTERVENTION CONSTRUCTION

C.1 CONSTRUCTION PIPELINE

Our minimal, label-changing revisions follow the local-control principle of counterfactually augmented data and contrast sets (Kaushik et al., 2020; Gardner et al., 2020). For every answerable question with annotated evidence, we apply the following deterministic-first pipeline.

1. **Evidence closure.** Start from the annotated evidence spans and add coreferent or explicitly dependent sentences needed to interpret the target claim. This prevents an edit from leaving a contradictory duplicate elsewhere in the local context.
2. **Atomic answer graph.** Normalize the gold answer into typed slots (entity, scalar with unit, signed direction, Boolean polarity, comparison, or bounded set). Link each slot to the smallest evidence statement that determines it.
3. **Type-preserving proposal.** Apply an operator from Table 5. The proposed value is selected to remain locally plausible, preserve grammar and units, and differ from every answer alias in the original document.

Table 5: Counterfactual edit operators. All edits are restricted to the evidence closure and preserve answer type.

Operator	Example transformation	Required validation
Entity	Method A $\rightarrow$ Method B	same entity class; no residual coreference
Numeric	$4.2 \rightarrow 2.4$ points	unit and precision preserved; table/text agree
Direction	improves $\rightarrow$ degrades	predicate antonym valid in local syntax
Polarity	supported $\rightarrow$ not supported	all scoped negation dependencies updated
Comparison	A $>$ B $\rightarrow$ B $>$ A	metric and comparison set unchanged
Bounded set	$\{$A,B$\} \rightarrow \{$A,C$\}$	cardinality and entity types preserved

4. **Dependent rewrite.** Propagate the edit only through the evidence closure. Section headings, citations, surrounding methods, distractor paragraphs, and the question are frozen.
5. **Automatic validation.** Require answer change, type validity, lexical-diff locality, no remaining original-support span, no newly introduced contradiction, and agreement between the deterministic answer transformer and a document-only verifier.
6. **Held-out validation.** Review the complete evaluation pair without revealing training method or model outputs. Reject rather than repair ambiguous pairs after model evaluation begins.

C.2 Control documents

The four modified document versions distinguish sensitivity to answer-critical evidence from sensitivity to changes that should leave the answer unchanged.

- D^{para}: paraphrase only the answer-critical evidence while preserving every typed answer slot. This tests semantic invariance.
- D^{irr}: insert paragraphs from another paper matched in length, section type, and broad topic but unrelated to q. This tests distraction and context length.
- D^{drop}: remove the full evidence closure and repair only dangling discourse markers. The correct behavior is to state that the document is insufficient.
- D^{cf}: edit the answer-critical slots and all dependent mentions so that the document uniquely supports a^{cf}.

The context tokenizer length is matched between D^{orig} and D^{cf} up to the small unavoidable difference introduced by type-preserving replacements. For the conditional bypass computation, answer aliases are canonicalized before comparing a counterfactual output with a^{orig}.

D Evaluation protocol

D.1 Deterministic answer scoring

Each primary item stores its answer type, canonical value, and allowed aliases. Text is Unicode-normalized, case-folded where appropriate, and stripped of citation markers before comparison. Numeric values are parsed with units and converted to the annotated unit; no tolerance is allowed unless the source itself reports rounding. Boolean, direction, and comparison answers use a fixed ontology. Set-valued answers require exact normalized set equality. Outputs that provide both the original and counterfactual answer without resolving the question are incorrect. This restricted atomic scorer is why headline Acc and CFAcc do not require an LLM judge.

D.2 Atomic claim segmentation

Claim-level evaluation asks whether a response is supported by its cited source, as in attribution and citation evaluation (Rashkin et al., 2023; Menick et al., 2022; Gao et al., 2023; Yue et al., 2023). Although the evaluation prompt requests one claim per step, we segment the output before assigning support labels. Coordinated propositions are split when either can be contradicted independently, and cited premises are separated from a derived conclusion. An equation, its substituted values, and

its verbal interpretation are treated as separate units when each can fail independently; discourse markers and restatements are omitted.

Segmentation uses anonymized outputs without method, checkpoint, document-version, or answer-correctness metadata. Boundary cases are resolved with a fixed guide before support labels are revealed. This procedure reduces the extent to which differences in verbosity or sentence structure change the unit being evaluated. These labels assess source support, not whether the claims describe the model's internal reasoning.

D.3 Structured evidence response prompt

```
You are answering a question using only the supplied scientific
document.
For each reasoning step, return exactly:
CLAIM: <one atomic claim>
EVIDENCE: <paragraph IDs that support this claim, or NONE>
RELATION: direct | composed | uncertain
After the steps, return:
ANSWER: <answer, or INSUFFICIENT EVIDENCE>
Do not use facts that are absent from the document.  If the
necessary evidence is missing, abstain.
```

All systems receive this response contract only at evaluation; it is not a method-specific prompt. Paragraph identifiers are assigned before document variants are generated so that unchanged paragraphs retain the same IDs.

E Paired-document component ablation

Configurations. We examine the contributions of counterfactual document exposure, answer supervision, and the paired margin on QASPER with the Qwen3-4B student. Each paired-document configuration uses original and counterfactual documents supporting different answers to the same question. Let

$$\mathcal{L}_{\text{dist}} = \frac{1}{2} \sum_{b \in \{0,1\}} \mathcal{L}_{\text{OPD}}(x^b; \theta)$$

denote the paired-document distillation term, and let $\mathcal{L}_{\text{CE}}$ denote cross-entropy supervision of each document's supported canonical answer. The compared objectives are

$$\begin{aligned} \mathcal{L}_{\text{Augmented OPD}} &= \mathcal{L}_{\text{dist}}, \\ \mathcal{L}_{\text{Paired ANS-SFT}} &= \mathcal{L}_{\text{CE}}, \\ \mathcal{L}_{\text{Augmented OPD+CE}} &= \mathcal{L}_{\text{dist}} + \gamma \mathcal{L}_{\text{CE}}, \\ \mathcal{L}_{\text{EP-OPD}} &= \mathcal{L}_{\text{dist}} + \lambda \mathcal{L}_{\text{pair}}, \end{aligned}$$

where γ weights answer cross-entropy. Augmented OPD removes the answer-level objective from EP-OPD; Paired ANS-SFT uses answer supervision without distillation; and Augmented OPD + Answer CE replaces the paired margin with cross-entropy. Standard OPD provides the original-document reference.

Document exposure and the paired margin. Table 6 shows that adding counterfactual documents changes the evidence-response profile: Augmented OPD reduces bypass from 35.8% to 27.2% and increases CSP from 57.0% to 62.3%, while CFAcc changes from 39.2% to 38.5%. Lower original-answer persistence thus accompanies a different change in counterfactual correctness, highlighting the complementary information provided by these metrics. Adding the paired margin increases CFAcc from 38.5% to 42.7%, raises CSP from 62.3% to 65.4%, and further reduces bypass to 25.9%, with original accuracy changing from 52.1% to 51.9%.

Answer supervision and objective choice. Paired ANS-SFT achieves the lowest bypass rate, 25.3%. Relative to this configuration, EP-OPD raises original accuracy by 1.4 percentage points, CFAcc by 0.7 points, and CSP by 1.6 points, with a 0.6-point increase in bypass. Combining

Table 6: Paired-document component ablation on QASPER with Qwen3-4B. Scores are percentages. Standard OPD uses original documents; the remaining configurations use original and counterfactual documents. Bold denotes the best value in each column.

Method	Orig.↑	CFAcc↑	Bypass↓	CSP↑
Standard OPD	52.4	39.2	35.8	57.0
Augmented OPD	52.1	38.5	27.2	62.3
Paired ANS-SFT	50.5	42.0	**25.3**	63.8
Augmented OPD + Answer CE	**52.5**	40.9	26.7	59.7
EP-OPD	51.9	**42.7**	25.9	**65.4**

Table 7: Document-operator and claim-level outcomes on QASPER–4B. Changes compare standard OPD with the frozen base. Scores are percentages, except the paraphrase gap and the change column, which are in pp. The expert row uses the blinded validation subset.

Outcome	Base	Standard OPD	OPD − Base
Original-document accuracy	49.0	52.4	+3.4
Paraphrase accuracy	48.7	51.9	+3.2
Paraphrase gap (Para. − Orig.)	−0.3	−0.5	−0.2
Counterfactual accuracy	44.5	39.2	−5.3
Evidence-drop abstention	22.0	14.4	−7.6
Conditional bypass	20.1	35.8	+15.7
Automatic step evidence error	30.2	43.0	+12.8
Expert step evidence error	28.5	38.3	+9.8

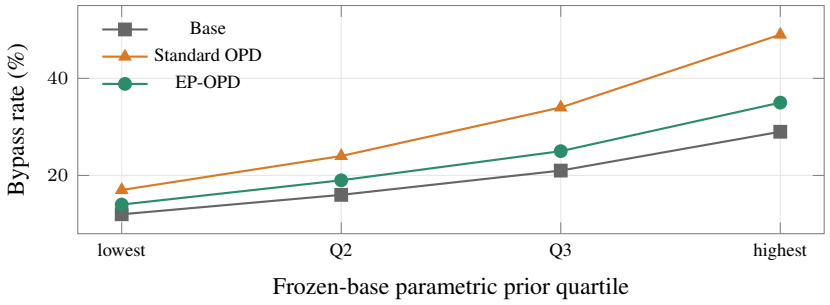

Figure 4: **Bypass concentrates on strong parametric priors.** Quartiles use the frozen base model's closed-book answer margin before any training. OPD's increment grows monotonically from +5 to +20 pp; EP-OPD attenuates it in every quartile.

augmented OPD with answer cross-entropy yields the highest original accuracy, 52.5%. Compared with this objective, EP-OPD improves CFAcc by 1.8 points and CSP by 5.7 points, while reducing bypass by 0.8 points at a 0.6-point original-accuracy cost. These comparisons characterize the joint contribution of on-policy supervision and document-specific answer ranking: the complete objective attains the highest CFAcc and CSP in the ablation while retaining an original accuracy of 51.9%.

F Additional Results

F.1 Document-Operator Localization

Answer-preserving paraphrase and irrelevant-insertion controls retain most of the observed-document gain. Evidence deletion and answer-changing interventions produce the largest changes in evidence-sensitive outcomes, and the claim-level rows show the same source-support pattern.

Table 8: Bypass rate (%) by frozen-base parametric prior quartile. Parentheses show the change from base.

Method	Lowest	Q2	Q3	Highest
BASE	12	16	21	29
Standard OPD	17 (+5)	24 (+8)	34 (+13)	49 (+20)
EP-OPD	14 (+2)	19 (+3)	25 (+4)	35 (+6)

Table 9: Implications and observed outcomes. Changes are relative to the frozen base unless a repair comparison is stated; main-setting results refer to QASPER–4B.

Implication	Expected direction	Observed
1. Observational competence rises	$\Delta^{\text{orig}}_{\text{OPD}} > 0$	$+3.4$ pp (main); $+2.1$ to $+3.4$ pp across settings
2. Evidence response separates	$\Psi_{\text{OPD}} < 0$	-8.7 pp (main); $\Delta\text{CFAcc} = -5.3$ pp
3. Original answer persists	$\Delta\text{BypassRate} > 0$	$+15.7$ pp (main)
4. Reallocation replicates and increases with prior strength	Same signs across settings; larger bypass increment at higher prior	Three other settings; quartile increments $+5$ to $+20$ pp
5. Pairing restores evidence response	$\text{CFAcc}_{\text{EP}} > \text{CFAcc}_{\text{OPD}}$	54–66% of CFAcc loss recovered
6. Repair preserves the original-task benefit	$\text{Acc}^{\text{EP}}_{\text{orig}} > \text{Acc}^{\text{Base}}_{\text{orig}}$	80–85% of gain retained; -0.5 pp vs. OPD (main)
7. Claim support tracks reallocation and repair	Evidence error rises after OPD; CSP rises with repair	Expert error $+9.8$ pp; repair CSP $+8.4$ pp (main)

F.2 PARAMETRIC-PRIOR STRATA

The base model already shows greater original-answer persistence in the stronger-prior quartiles. Our main comparison concerns the additional increase after OPD, which grows from $+5$ to $+20$ pp across the quartiles. EP-OPD reduces most of this training-induced increase, while the differences already present in the base model remain.

F.3 EMPIRICAL DIAGNOSTICS

These comparisons bring together the main finding, its replication, and the effect of evidence pairing. The first four describe how answer accuracy and evidence response change after OPD, including their relationship to the student's initial answer preference. The remaining comparisons assess whether EP-OPD improves evidence response and claim support while retaining the original-task gain. Together with the document controls in Table 7, they help assess alternatives such as general perturbation sensitivity or indiscriminate answer suppression. The direction column summarizes the qualitative relationships under examination; the observed column gives the reported magnitudes.

G ILLUSTRATIVE INTERVENTION EXAMPLES

Direction edit. "The proposed regularizer *improves* calibration" becomes "*worsens* calibration." The question and experiment description are unchanged; the correct comparison reverses.

Numeric edit. "Ablation reduces F1 by 4.2 points" becomes "0.4 points." Units, table position, and qualitative framing are preserved.

Evidence deletion. Sentences locating an out-of-domain split are removed while dataset names remain in unrelated paragraphs. A supported system must abstain.

These schematic examples use synthetic text for illustration. The manifest records each exact diff, answer transformation, evidence closure, and validation status.