DOMAIN-GROUNDED COUNTERFACTUAL AUGMENTATION FOR MITIGATING KNOWLEDGE CONFLICTS IN LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain-specific fine-tuning of pre-trained LLMs exposes a persistent failure mode: existing parametric knowledge persists as a dominant inductive bias, so models produce confident, internally consistent predictions that remain inconsistent with the target domain even after supervised fine-tuning on it. Such failures are not fully captured by label noise or distributional shift in the conventional sense; they reflect a domain-conditioned knowledge-priority problem, in which generalizations valid under the original training distribution are overridden by more specific target-domain facts, which creates a need for targeted supervision and consistency across domain-preserving views. We introduce **CIDA** (Conflict-Immunized Data Augmentation), a framework that applies counterfactual invariance to this misalignment. CIDA identifies samples prone to conflict via distributional consistency under stochastic decoding, constructs knowledge-conditioned rewrites intended to vary prior-triggering surface cues while preserving the target answer, and optimizes a composite objective coupling conflict-aware sample weighting with an invariance regularizer on predictions over original and counterfactual pairs. Evaluated on an internal domain-specific dataset and public benchmarks covering diverse domains, CIDA outperforms competitive baselines from data filtering and sample reweighting paradigms. Gains concentrate on high-conflict samples, while low-conflict performance remains comparable to standard fine-tuning. Sample- and update-matched comparisons separate additional training exposure from the augmentation pipeline, while a same-data ablation identifies the additional benefit of paired invariance. Together, these comparisons support retrieval-grounded augmentation and prediction consistency as complementary contributors to the gains.

1 INTRODUCTION

Domain-specific fine-tuning of large language models (Brown et al., 2020; Touvron et al., 2023) often encounters *knowledge conflicts*: the model's parametric knowledge contradicts domain facts, producing internally consistent but domain-inconsistent outputs that persist after standard supervised fine-tuning. Such conflicts arise whenever domain-specific facts systematically contradict the model's existing priors (Longpre et al., 2021; Xu et al., 2024), for instance in medical protocols that override general clinical knowledge or legal regulations that depart from common-law principles (Feng et al., 2026; Wang et al., 2024). Unlike hallucination (a symptom) or catastrophic forgetting (knowledge loss), knowledge conflict occurs when old and new knowledge *coexist but contradict*, making the model's behavior unreliable precisely on the samples where domain adaptation matters most. Figure 1 illustrates this phenomenon and previews the CIDA remedy.

We approach this problem through an intervention perspective (Pearl, 2009; Peters et al., 2017): vary question phrasing that can trigger an inappropriate prior while preserving the intended domain answer. Filtering and reweighting (Zhong et al., 2025; Lyu et al., 2024) alter sample selection or emphasis, retrieval augmentation (Lewis et al., 2020; Izacard et al., 2023) supplies relevant evidence, and knowledge editing (Meng et al., 2022; 2023) directly updates model parameters. CIDA uses

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.

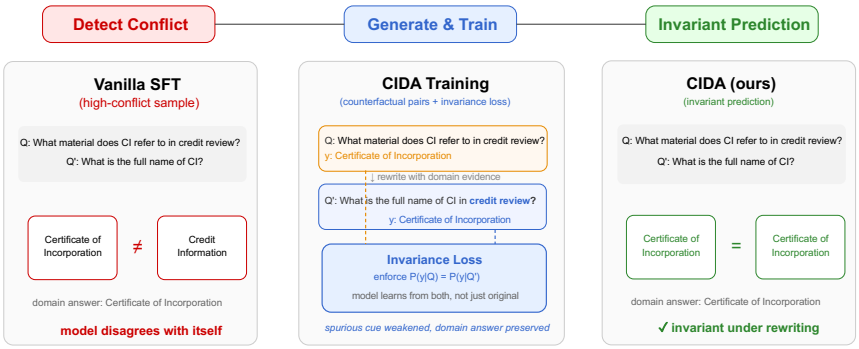


Figure 1: Schematic illustration of domain-conditioned question rewriting. The intended setting throughout is credit review. Left: an abbreviation receives different interpretations when the domain context is implicit. Middle: the rewrite makes that context explicit while retaining the intended answer. Right: paired training promotes consistent domain-specific answers.

retrieved evidence to construct paired training views and couples their predictions. This design lets us test whether targeted rewriting and paired training improve adaptation beyond the additional answer-supervised examples.

Our key insight is that conflicts can be mitigated by teaching the model to produce *invariant* predictions under counterfactual perturbations that weaken spurious cues while preserving domain semantics. We generate answer-preserving rewrites that replace generic phrasing triggering the model's priors with domain-specific language anchored in target-domain evidence, and enforce consistency between original and counterfactual pairs via an invariance regularizer. This design follows the intervention perspective of counterfactual augmentation (Kaushik et al., 2020): supervised answers establish the desired prediction, while paired consistency discourages dependence on the particular surface cues changed by the rewrite.

We introduce **CIDA** (**C**onflict-**I**mmunized **D**ata **A**ugmentation), a four-stage framework: (1) annotation-free conflict detection via multi-temperature sampling consistency, (2) knowledge-conditioned counterfactual generation using retrieved domain evidence, (3) multi-level quality filtering, and (4) conflict-immunized training with invariance regularization. The training objective combines conflict-aware weighted cross-entropy with a counterfactual invariance loss that enforces prediction consistency between original and counterfactual inputs via KL divergence. A motivating analysis relates paired prediction agreement to bounded task loss and states the local gradient-alignment condition under which consistency can improve task risk. Across four benchmarks (MedMCQA, PopQA, SciQ, and a proprietary credit-risk dataset CreditQA), CIDA consistently outperforms six baselines, with gains of $+12.10$ points on PopQA and $+8.11$ points on SciQ over Vanilla FT; ablations identify counterfactual generation as the single most critical component, and additional experiments characterize the pipeline on multi-hop and temporal questions.

Our contributions are:

- A model-relative formulation of stable disagreement with domain answers that guides selective augmentation and distinguishes the targeted failure pattern from general prediction uncertainty (Section 3.1).
- The CIDA framework: annotation-free conflict detection, knowledge-conditioned counterfactual generation with four-level quality filtering, and conflict-immunized training with composite invariance regularization (Section 3).
- A motivating analysis of prediction stability on answer-preserving pairs and the local conditions for beneficial consistency updates, with matching derivations (Section 4).
- Empirical validation across four benchmarks and three model families, including sample- and update-matched augmentation controls, a same-data ablation of paired invariance, and an answer-recoverability audit (Section 5). Additional experiments characterize the pipeline's behavior on multi-hop and temporal questions.

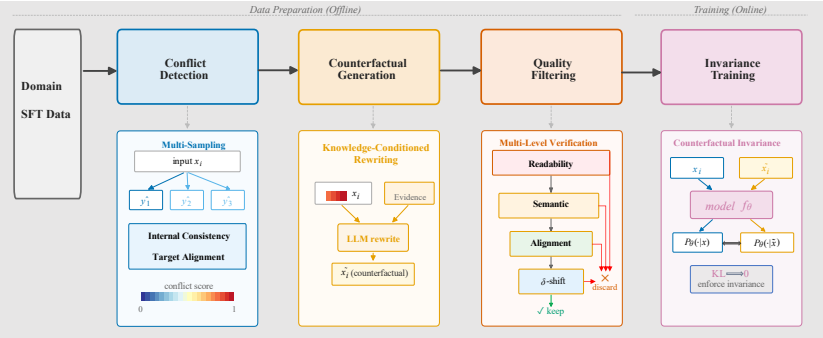


Figure 2: Overview of the CIDA framework. Stage 1 detects conflicts via multi-temperature sampling. Stage 2 generates knowledge-conditioned counterfactuals. Stage 3 filters with four-level quality verification. Stage 4 trains with conflict-immunized composite loss.

2 Related Work

Knowledge conflicts arise when parametric knowledge contradicts external evidence or domain facts (Longpre et al., 2021; Xu et al., 2024). Prior work studies how LLMs navigate them (Feng et al., 2026; Wang et al., 2024) and shows models struggle with long-tail knowledge (Mallen et al., 2023). Fine-tuning on conflicting data does not reliably fix this: Kazemi et al. (2023) find fine-tuning mainly teaches QA format rather than new knowledge, and Ghosal et al. (2024) identify a "frequency shock" on less-known facts. Existing remedies—data filtering, sample reweighting (Zhong et al., 2025; Lyu et al., 2024), retrieval augmentation (Lewis et al., 2020; Izacard et al., 2023), and knowledge editing (Meng et al., 2022; 2023)—address sample coverage, training emphasis, evidence access, or parameter updates. CIDA combines conflict-targeted rewriting with a consistency penalty that links original questions to their domain-grounded views.

Counterfactual data augmentation has been used to reduce spurious correlations in NLU tasks (Kaushik et al., 2020) and, more recently, to systematically generate counterfactuals with LLMs (Nguyen et al., 2024; Steindl et al., 2024). Two lines of work apply counterfactuals to knowledge conflict most directly. Neeman et al. (2023) train a QA model to disentangle parametric and contextual knowledge, using counterfactual data augmentation so that the model can emit separate contextual and parametric answers at inference time. Song et al. (2024) use controllable counterfactual data augmentation to improve the entity-level factual adaptiveness of fine-tuning-based summarization models. CIDA differs from both in what the counterfactual controls: rather than disentangling knowledge sources at inference time (DisentQA) or tuning entity-level factual signals for summarization (Song et al.), we use a conflict-score-driven, training-time intervention plus an invariance regularizer on paired original–counterfactual inputs, to reduce prediction sensitivity to the surface cues varied by the rewrite. Our approach is further distinguished by conditioning counterfactual generation on retrieved domain evidence, guiding rewrites toward domain terminology and facts. Our design draws from invariant learning (Arjovsky et al., 2020; Krueger et al., 2021; Lai & Wang, 2024), but rather than requiring multiple environments we construct them via counterfactual intervention—an original and its counterfactual are two views of the same content—enforcing invariance without multi-domain data. The regularizer is a function-space constraint on prediction consistency, unlike parameter-space regularizers such as EWC (Kirkpatrick et al., 2017) or LoRA (Hu et al., 2022). Extended discussion is in Appendix B.

3 Methodology

Having characterized knowledge conflict as a mismatch between pre-trained priors and domain supervision, we now describe CIDA, which turns that mismatch into a targeted training signal. The method identifies stable disagreement, rewrites the affected questions using domain evidence, and couples supervision on the original and rewritten views with a prediction-consistency objective. Figure 2 summarizes the four stages.

3.1 Problem Formulation

We begin by formalizing the disagreement that determines where CIDA intervenes. Consider a pre-trained LLM $\mathcal{M}_\theta$ and a domain-specific SFT dataset $\mathcal{D}_{\text{sft}} = \{(x_i, y_i)\}_{i=1}^{N}$. Standard fine-tuning minimizes cross-entropy uniformly over the samples, implicitly treating the model's existing features as equally reliable throughout the dataset. This assumption breaks down for *knowledge conflicts*, where the model's parametric knowledge contradicts the domain-specific answer y_i.

We operationalize a conflict as stable disagreement with the domain answer. Diverse answers across decoding configurations indicate uncertainty, while mutually consistent answers that disagree with y_i receive higher priority for intervention. The score combines internal consistency and target alignment:

$$s_i = \begin{cases} 1 - \text{Sim}(\hat{y}_0, y_i) & \text{if } \text{Cons}(\hat{y}_1, \ldots, \hat{y}_K) > \tau_c, \\ \frac{1}{2}\big(1 - \text{Sim}(\hat{y}_0, y_i)\big) & \text{otherwise,} \end{cases} \tag{1}$$

where $\hat{y}_1, \ldots, \hat{y}_K$ denote the K decoded outputs and $\hat{y}_0$ denotes the greedy member of this collection. Thus the greedy reference is included in the output budget rather than an additional sample. When outputs are mutually consistent but disagree with the target ($\text{Cons} > \tau_c$), the full misalignment score is assigned; when the model is uncertain ($\text{Cons} \leq \tau_c$), the score is halved to discount uncertainty while retaining a signal of target disagreement. Clipping $\text{Sim}(\hat{y}_0, y_i) \in [0, 1]$ gives $s_i \in [0, 1]$, keeping the original weight $w_{\text{orig},i} = 1 - \alpha_p s_i$ strictly positive for any $\alpha_p \in (0, 1)$ (cf. Stage 4). Given (x, y) and retrieved knowledge k, a valid counterfactual $\tilde{x} = \text{CF}(x, k)$ preserves the intended question and its answer. Reducing the original conflict is a design goal, assessed separately from semantic preservation.

3.2 Stage 1: Annotation-Free Conflict Detection

With the target failure pattern defined, the first stage determines where intervention is needed without collecting separate conflict annotations. It reuses the task labels y_i in $\mathcal{D}_{\text{sft}}$ to measure target alignment; the detector is therefore conflict-annotation-free rather than fully unsupervised. The key signal is that a strong but incorrect prior tends to produce outputs that remain mutually consistent across generation configurations while disagreeing with the domain answer.

For each x_i we sample K outputs at multiple temperatures: the lowest-temperature output is the baseline $\hat{y}_0$, and higher-temperature outputs reveal whether the model has a stable prior. Internal consistency is the average pairwise cosine similarity of sentence embeddings (Reimers & Gurevych, 2019) across outputs, and alignment is the similarity between $\hat{y}_0$ and the target. High consistency with low alignment identifies stable target disagreement. The score is an operational selection signal, with interpretation and threshold sensitivity discussed in Appendix G.5.

The score drives two mechanisms. It selects which samples receive counterfactual augmentation in Stage 2 (only $s_i \geq \tau_{\text{cf}}$), and it modulates Stage 4 weights: high-conflict originals are down-weighted ($w_{\text{orig}} = 1 - \alpha_p s_i$) while their counterfactuals are up-weighted ($w_{\text{cf}} = 1 + \alpha_r s_i$). This asymmetry reduces exposure to conflicting originals and increases exposure to domain-consistent counterfactuals, shifting supervised emphasis toward the grounded views.

3.3 Stage 2: Knowledge-Conditioned Counterfactual Generation

Detection tells us which questions are problematic, but not how to make them more informative for training. The second stage therefore constructs counterfactuals that vary prior-triggering surface cues while retaining the information that determines the domain answer. Because a generic paraphrase may preserve or even introduce spurious cues, the rewrite is conditioned on retrieved domain evidence and grounded in target-domain facts.

For each sample selected by $s_i \geq \tau_{\text{cf}}$, we first retrieve relevant domain evidence from a FAISS index built from the training corpus using sentence-transformer embeddings (Reimers & Gurevych, 2019). We then classify the conflict type (knowledge, semantic, or temporal) based on linguistic cues in the question, and use the LLM itself to rewrite the question conditioned on the retrieved evidence and conflict type:

$$\tilde{x}_i = \mathcal{M}_\theta\big(\text{Prompt}_{\text{CF}}(x_i, y_i, \{e_j\}_{j=1}^{k}, t_i)\big) \tag{2}$$

The prompt instructs the model to replace generic language that triggers its priors with domain-specific terminology anchored in the evidence, remove ambiguity, and sharpen the intended answer without revealing it. Because $\mathcal{M}_\theta$ itself generates $\tilde{x}_i$, its spurious prior could leak into the rewrite; we mitigate this in two ways: the retrieved evidence $\{e_j\}$ anchors the rewrite toward domain terminology, and the Level-3 knowledge-alignment filter (Stage 3) rejects rewrites whose alignment falls below τ_{align}. A dedicated leakage audit appears in Section 5.4 and Table A6.

Training supervision and retrieval scope. The target y_i is available during training and guides answer-preserving generation. Retrieval uses the training corpus, so the retrieved material is an additional source of training supervision rather than independent evidence of absence of answer cues. The audit in Section 5.4 measures answer information retained in the rewritten input, and the same-data ablation holds that information fixed when testing the invariance term. The present comparison concerns answer-conditioned augmentation. An answer-blind generator would change the supervision available to the augmentation stage and constitutes a separate control.

3.4 Stage 3: Multi-Level Quality Filtering

Grounded generation does not by itself guarantee a usable training pair: a rewrite may still add noise, change meaning, or leave the relevant cues untouched. The third stage therefore applies a progressive four-level filter. Level 1 checks *readability*: length within bounds, without excessive special characters or repetition. Level 2 checks *semantic preservation* through $\text{Sim}(x_i, \tilde{x}_i) > \tau_{\text{sem}}$. Level 3 checks *knowledge alignment*, requiring sufficient similarity between $\tilde{x}_i$ and the retrieved evidence. Level 4 enforces *bounded semantic change*: $\delta = 1 - \text{Sim}(x_i, \tilde{x}_i)$ must lie in $[\delta_{\min}, \delta_{\max}]$. This proxy selects a nontrivial but bounded change in phrasing; any decrease in the model-based conflict score must be assessed separately.

Counterfactuals passing all levels are retained. Rejected rewrites are discarded while their originals remain available. The Stage 4 counterfactual weight uses the *original* score s_i, so rewrites of the most conflicted questions receive the highest weight. When conflict detection is repeated on a rewrite, $\tilde{s}_i$ denotes the output of the same scoring rule and $\Delta s_i = s_i - \tilde{s}_i$ is a signed diagnostic. Its positivity is not implied by the semantic-distance filter. Appendix F describes the hybrid detector check.

3.5 Stage 4: Conflict-Immunized Training

After the pairs have been selected and validated, the final stage turns them into a joint training signal. Its composite objective contains two supervised terms and one consistency term, allowing CIDA to change both the emphasis of supervision and the relation learned between paired inputs.

Conflict-aware weighted cross-entropy. The supervised terms determine which examples receive the greatest emphasis. Because some original questions reinforce the prior against the domain answer, the objective down-weights them by $w_{\text{orig},i} = 1 - \alpha_p s_i$ and up-weights their counterfactuals by $w_{\text{cf},i} = 1 + \alpha_r s_i$:

$$\begin{aligned} \mathcal{L}_{\text{orig}} &= \frac{1}{N} \sum_{i=1}^{N} (1 - \alpha_p \cdot s_i) \cdot \ell_i, \\ \mathcal{L}_{\text{cf}} &= \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} (1 + \alpha_r \cdot s_i) \cdot \ell_i^{\text{cf}} \end{aligned} \tag{3}$$

where ℓ_i and ℓ_i^{cf} are cross-entropy losses on the shared target answer under original and counterfactual inputs, and the sum over $\mathcal{C}$ runs over retained original–rewrite pairs, counting each retained rewrite separately when an original has multiple rewrites. Both branches receive direct supervision from the domain answer. The weighting shifts emphasis toward domain-grounded rewrites while retaining supervision on the original questions, which are the inputs of interest at evaluation.

Counterfactual invariance loss. Weighting alone does not prevent the model from relying on spurious features that happen to correlate with the correct answer in training. The third term therefore

penalizes prediction disagreement between paired inputs:

$$\mathcal{L}_{\text{inv}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \text{KL}\big(P_\theta(\cdot \mid x_i) \,\big\|\, P_\theta(\cdot \mid \tilde{x}_i)\big) \tag{4}$$

Gradients flow through both branches, so neither prediction is a fixed teacher. The KL term penalizes disagreement, while the two supervised losses determine the desired answer. Its role is to couple the learning of the original question and its grounded rewrite, encouraging their predictions to agree as both learn the domain target. The same-data ablation in Section 5.3 evaluates this coupling separately from the additional supervised examples. Here $P_\theta(\cdot \mid x)$ denotes an answer distribution: the equation is a distribution-level consistency objective, not a comparison of the probabilities assigned only to the gold answer. Appendix G.3 distinguishes complete-sequence KL from a teacher-forced answer-token surrogate, including the shared-prefix alignment and vocabulary over which token distributions are compared.

Complete training objective. The total loss combines all three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{orig}} + \mathcal{L}_{\text{cf}} + \lambda_{\text{inv}} \mathcal{L}_{\text{inv}} \tag{5}$$

where λ_{inv} balances the invariance regularizer against the cross-entropy terms. Sample weights vary continuously with s_i: low-conflict originals retain weights close to one, while more conflicted originals receive lower weights and their retained counterfactuals receive higher weights. The low/mid/high bins used in reporting summarize this continuous score and do not replace the weighting equations.

4 Motivating Analysis

The preceding section specifies how CIDA combines supervision on grounded rewrites with a penalty on prediction disagreement. We now examine what that penalty can—and cannot—justify. The analysis separates prediction stability across valid answer-preserving pairs from the local effect of a consistency update on task risk; it motivates the objective without claiming identification of latent causal features. Appendix G provides the derivations and the autoregressive interpretation of the divergence.

Prediction stability on paired inputs. We first ask how much the expected prediction loss can change between two valid views of the same question. Let Q be a distribution over answer-preserving triples $(X, \tilde{X}, Y)$, and let $P_\theta(\cdot \mid x)$ denote a distribution over answers. For any loss $\ell(a, y) \in [0, 1]$, define $R_x(\theta) = \mathbb{E}_Q \mathbb{E}_{A \sim P_\theta(\cdot|X)}[\ell(A, Y)]$ and define $R_{\text{cf}}(\theta)$ analogously using $\tilde{X}$. Let $\mathcal{G}_{\text{inv}}(\theta) = \mathbb{E}_Q[\text{KL}(P_\theta(\cdot \mid X) \| P_\theta(\cdot \mid \tilde{X}))]$ be finite.

Proposition 1 (Paired prediction stability)**.** *For the same pair distribution Q and bounded loss defined above,*

$$\big|R_x(\theta) - R_{\text{cf}}(\theta)\big| \le \sqrt{\mathcal{G}_{\text{inv}}(\theta)/2}. \tag{6}$$

This follows from Pinsker's inequality and Jensen's inequality. The result concerns the expected loss of a prediction drawn from the model distribution. It explains why accurate grounded predictions and low disagreement are complementary: supervision improves the answer distribution, and consistency limits its change across paired inputs. Agreement alone does not establish correctness. The benchmark accuracy of decoded answers is evaluated separately in Section 5.

When consistency helps the task objective. The stability bound explains the consequence of low disagreement, but it does not imply that optimizing consistency will reduce task error. To make that distinction explicit, let R_H be a differentiable surrogate risk on a fixed high-conflict subset. At parameters θ, write $g_{\text{sup}} = \nabla(\mathcal{L}_{\text{orig}} + \mathcal{L}_{\text{cf}})(\theta)$ and $g_{\text{inv}} = \nabla \mathcal{L}_{\text{inv}}(\theta)$. We compare a supervised step $\theta_b = \theta - \eta g_{\text{sup}}$ with a paired step $\theta_p = \theta_b - \eta \lambda_{\text{inv}} g_{\text{inv}}$, for $\eta, \lambda_{\text{inv}} > 0$.

Proposition 2 (Local alignment condition)**.** *If R_H has an L-Lipschitz gradient on the segment joining θ_b and θ_p, then*

$$\begin{aligned} R_H(\theta_p) - R_H(\theta_b) \le &-\eta \lambda_{\text{inv}} \langle \nabla R_H(\theta_b), g_{\text{inv}} \rangle \\ &+ \frac{L \eta^2 \lambda_{\text{inv}}^2}{2} \|g_{\text{inv}}\|^2. \end{aligned} \tag{7}$$

A sufficient condition for improvement is that the inner product exceeds $(L\eta\lambda_{\text{inv}}/2)\|g_{\text{inv}}\|^2$. Its sign is a condition to be checked, not a consequence of low KL divergence. This local Euclidean-step comparison motivates preserving answer semantics and pairing related questions. The same-data ablation in Section 5.3 measures the actual training benefit, while random pairing tests the effect of disrupting that correspondence.

5 EXPERIMENTS

We evaluate CIDA through four questions: overall accuracy, the contribution of each mechanism, behavior across conflict levels, and possible answer leakage.

5.1 EXPERIMENTAL SETUP

We use a common protocol across datasets, with each control changing only the data or objective component it is intended to test.

Datasets. We evaluate on four benchmarks spanning different conflict types: **CreditQA** (6,937 QA pairs from a real-world lending domain, proprietary), **MedMCQA** (Pal et al., 2022) (medical, 182K questions), **PopQA** (Mallen et al., 2023) (long-tail factual knowledge, 14K questions), and **SciQ** (Welbl et al., 2017) (science domain multiple choice, 13K questions). All four datasets share a key property: the correct answer is a single fact that should remain unchanged under counterfactual rewriting of the question, which is the setting CIDA is designed for. After conflict detection and counterfactual augmentation, training sets contain 9,662 (CreditQA), 18,383 (MedMCQA), 18,130 (PopQA), and 15,183 (SciQ) samples. Detailed statistics are in Appendix C.1.

Baselines and metrics. We compare against six baselines: Vanilla FT (standard fine-tuning), Drop-high (remove high-conflict samples), Weighted-SFT (uniform conflict-based reweighting), KaFT (Zhong et al., 2025) (learned consistency-based reweighting), FT-Top (top-$k\%$ low-conflict selection), and KnowTuning-Lite (Lyu et al., 2024) (perplexity-based hard-sample enhancement). Full KnowTuning additionally requires a reference model and an external domain knowledge base for fine-grained augmentation. Neither is available for proprietary CreditQA, so we use the Lite variant consistently across all four datasets (Appendix B).

We report overall accuracy, accuracy in fixed conflict-score bins (Acc_{low} / Acc_{mid} / Acc_{high}), and original–counterfactual prediction consistency. The full comparison and ablations use Qwen3-8B (Yang et al., 2025); on Llama-3-8B (Grattafiori et al., 2024) and Mistral-7B (Jiang et al., 2023), we compare CIDA with Vanilla FT, Drop-high, FT-Top, and KnowTuning-Lite. All models use LoRA, and results are mean $\pm$ standard deviation over three seeds. We interpret differences in means descriptively rather than assigning significance from these summaries. Table 2 includes an original-budget Vanilla SFT reference and controls matched to CIDA in sample count and optimization steps; other hyperparameters remain fixed unless explicitly ablated. Details appear in Appendices C.4 and C.3.

5.2 MAIN RESULTS

We first compare aggregate accuracy across domains and model families.

CIDA attains the highest mean accuracy among the compared methods on all four benchmarks (Table 1). On Qwen3-8B, its gains over Vanilla FT are 7.63 points on CreditQA, 6.73 on MedMCQA, 12.10 on PopQA, and 8.11 on SciQ. Relative to the strongest baseline for each dataset, CIDA gains 5.86 points on PopQA and 4.27 on SciQ over FT-Top, and 1.03 on CreditQA and 1.53 on MedMCQA over KnowTuning-Lite. The advantage also holds for Llama-3-8B and Mistral-7B; Section 5.4 locates it within each dataset.

PopQA combines the largest gain over Vanilla FT with the highest training-set conflict prevalence: 53% of its original questions have $s_i \geq 0.7$, compared with 32% for CreditQA, 30% for SciQ, and 17% for MedMCQA (Appendix C.2). These are training-pool proportions; the stratified evaluation below instead measures performance within evaluation bins.

Backbone	Method	Credit	Med	Pop	SciQ
Qwen3-8B	Vanilla FT	66.18 ± 0.72	52.14 ± 0.85	39.22 ± 1.02	74.47 ± 0.64
	Drop-high	69.47 ± 0.75	53.53 ± 0.79	42.08 ± 0.88	77.02 ± 0.67
	Weighted-SFT	68.32 ± 0.59	52.97 ± 0.90	41.05 ± 0.85	75.76 ± 0.63
	KaFT	70.54 ± 0.52	54.48 ± 0.68	43.97 ± 0.85	77.53 ± 0.54
	FT-Top	71.17 ± 0.63	54.82 ± 0.76	45.46 ± 0.72	78.31 ± 0.60
	KnowTuning-Lite	72.78 ± 0.49	57.34 ± 0.70	44.52 ± 0.77	77.19 ± 0.57
	CIDA (Ours)	73.81 ± 0.53	58.87 ± 0.61	51.32 ± 0.78	82.58 ± 0.47
Llama-3-8B	Vanilla FT	63.92 ± 0.78	49.87 ± 0.92	37.14 ± 1.03	71.95 ± 0.69
	Drop-high	65.73 ± 0.81	51.26 ± 0.88	39.57 ± 0.97	73.82 ± 0.71
	FT-Top	67.45 ± 0.76	52.84 ± 0.83	42.13 ± 0.91	75.67 ± 0.66
	KnowTuning-Lite	69.27 ± 0.70	54.56 ± 0.75	44.89 ± 0.85	77.54 ± 0.60
	CIDA (Ours)	70.94 ± 0.82	55.83 ± 0.68	47.28 ± 0.79	79.47 ± 0.52
Mistral-7B	Vanilla FT	61.85 ± 0.84	48.06 ± 0.98	34.89 ± 1.12	69.92 ± 0.75
	Drop-high	63.64 ± 0.87	49.52 ± 0.95	37.25 ± 1.06	71.78 ± 0.78
	FT-Top	65.37 ± 0.80	51.08 ± 0.89	39.86 ± 1.01	73.63 ± 0.72
	KnowTuning-Lite	67.12 ± 0.69	52.79 ± 0.81	42.67 ± 0.92	75.41 ± 0.65
	CIDA (Ours)	68.29 ± 0.63	53.97 ± 0.74	44.96 ± 0.86	76.98 ± 0.58

Table 1: Main results across four benchmarks. Entries are accuracy (%, mean $\pm$ SD over three seeds).

Variant	Credit	Med	Pop	SciQ	Avg.	Δ vs Full CIDA
Full CIDA	73.81 ± 0.53	58.87 ± 0.61	51.32 ± 0.78	82.58 ± 0.47	66.65 ± 0.38	–
w/o Quality Filtering	72.21 ± 0.57	57.29 ± 0.67	49.48 ± 0.76	81.42 ± 0.51	65.10 ± 0.37	-1.55
w/o Conflict Detection	71.58 ± 0.59	56.83 ± 0.70	48.77 ± 0.79	80.82 ± 0.53	64.50 ± 0.39	-2.15
w/o Invariance Loss	71.42 ± 0.61	56.48 ± 0.72	48.21 ± 0.81	80.29 ± 0.54	64.10 ± 0.40	-2.55
w/o Retrieval Grounding	70.97 ± 0.63	56.23 ± 0.75	47.78 ± 0.84	79.81 ± 0.56	63.70 ± 0.42	-2.95
$\star$ Invariance on Random Pairs	69.74 ± 0.68	55.16 ± 0.79	45.32 ± 0.88	78.07 ± 0.60	62.07 ± 0.44	-4.58
w/o CF Generation	67.38 ± 0.71	54.42 ± 0.83	44.47 ± 0.92	77.03 ± 0.62	60.83 ± 0.47	-5.82
$\star$ Generic Paraphrase SFT	68.36 ± 0.73	54.05 ± 0.86	43.18 ± 0.95	76.94 ± 0.64	60.63 ± 0.49	-6.02
$\star$ Update-matched Vanilla SFT	67.05 ± 0.75	53.12 ± 0.89	41.27 ± 0.98	75.68 ± 0.66	59.28 ± 0.51	-7.37
Vanilla SFT only	66.18 ± 0.72	52.14 ± 0.85	39.22 ± 1.02	74.47 ± 0.64	58.00 ± 0.51	-8.65

Table 2: Ablations and matched controls. Entries are accuracy (%, mean $\pm$ SD over three seeds). Avg. is the unweighted average across the four datasets, and Δ is its change relative to Full CIDA. All rows except the final original-budget Vanilla SFT reference match CIDA's training sample count and optimization steps, with shared hyperparameters and seeds. $\star$ marks alternative data-construction or pairing controls.

The cross-backbone comparison provides a complementary check on the scope of this result. Absolute accuracy changes with the backbone, but the ordering between CIDA and the shared baselines is preserved across all four datasets. This consistency suggests that the benefit is not tied to one model's pre-training distribution or decoding behavior. At the same time, Table 1 should not be read as establishing architecture-invariant effect sizes: the three backbones may encode different conflicts, and the broader ablation suite is run on Qwen3-8B. We therefore treat the additional backbones as a transfer check and rely on the matched Qwen3-8B comparisons for mechanism attribution.

5.3 Mechanism-Isolation Ablations

Aggregate gains do not distinguish better data from additional exposure or paired consistency. We therefore use controls that hold the training budget or augmented data fixed while changing one mechanism at a time.

Update matching raises Vanilla SFT from 58.00% to 59.28%, still 7.37 points below Full CIDA (Table 2). Removing retrieval grounding yields 63.70%, while generic-paraphrase SFT obtains 60.63%; the former tests grounding within CIDA, whereas the latter compares complete augmentation pipelines.

Keeping the grounded data and weights but removing invariance yields 64.10%, versus 66.65% for Full CIDA, identifying a 2.55-point same-data increment from paired consistency. Random pairing obtains 62.07% and tests broken correspondence rather than data quality. Removing generation, detection, or filtering lowers accuracy by 5.82, 2.15, or 1.55 points, respectively; these conditional effects are not additive. Appendix C.4 specifies each comparison.

Dataset	Method	Low (< 0.3)	Mid (0.3–0.7)	High (≥ 0.7)	Cons.
CreditQA	Vanilla FT	95.17 ± 0.32	70.84 ± 0.87	54.26 ± 1.24	58.39 ± 1.08
	CIDA	95.03 ± 0.35	74.08 ± 0.79	71.54 ± 1.05	79.62 ± 0.96
	Δ	-0.14	$+3.24$	$+17.28$	$+21.23$
MedMCQA	Vanilla FT	75.06 ± 0.65	47.28 ± 1.02	22.01 ± 1.38	62.14 ± 1.12
	CIDA	75.09 ± 0.62	55.02 ± 0.88	39.57 ± 1.72	75.48 ± 0.99
	Δ	$+0.03$	$+7.74$	$+17.56$	$+13.34$
PopQA	Vanilla FT	65.04 ± 0.78	45.07 ± 1.15	32.53 ± 1.36	50.18 ± 1.27
	CIDA	65.58 ± 0.83	55.01 ± 0.94	47.09 ± 1.52	72.82 ± 1.05
	Δ	$+0.54$	$+9.94$	$+14.56$	$+22.64$
SciQ	Vanilla FT	91.58 ± 0.41	77.03 ± 0.76	60.05 ± 1.18	61.47 ± 1.02
	CIDA	91.82 ± 0.39	84.09 ± 0.81	74.51 ± 1.07	78.96 ± 0.89
	Δ	$+0.24$	$+7.06$	$+14.46$	$+17.49$

Table 3: Conflict-stratified accuracy and original–counterfactual prediction consistency (Cons.; %, mean $\pm$ SD over three seeds). Cons. measures prediction *stability* across an original and its counterfactual—whether the two answers agree—independent of whether the shared answer is correct. The Δ rows report CIDA minus Vanilla FT.

5.4 Conflict-Stratified, Consistency, and Leakage Analysis

We next test the central behavioral prediction: gains should concentrate at higher conflict scores without sacrificing low-conflict performance.

The results follow this pattern (Table 3): gains on high-conflict questions range from 14.46 to 17.56 points, whereas low-conflict changes remain between -0.14 and $+0.54$ points. Original–rewrite consistency rises by 13.34–22.64 points; because this metric includes agreement on incorrect answers, the accompanying accuracy gains are essential to its interpretation. Training-set score proportions are reported separately in Appendix C.2. Reading the stratified accuracy and consistency columns together helps distinguish useful robustness from indiscriminate agreement. If consistency increased without accuracy, the model might simply have learned to repeat the same wrong answer across both views. Instead, the largest consistency changes accompany substantial high-conflict accuracy gains, while low-conflict accuracy remains nearly unchanged. This pattern is consistent with reduced sensitivity to the cues altered by grounded rewriting rather than a global flattening of the predictions. It remains an empirical interpretation: the bins are defined by a model-relative score, and neither bin membership nor paired agreement identifies a causal feature. The matched controls and leakage audit are therefore still needed to support the mechanism claim.

Leakage and joint-control audit. Answer-conditioned generation raises a separate concern: a rewrite might reveal additional target cues. We therefore audit leakage directly. A held-out recovery probe yields AUC 0.53 ± 0.04 and a 0.71 ± 0.52-point increase over original questions; filtering literal overlap changes downstream accuracy by $+0.24 \pm 0.29$ points. Together with the matched controls, these results support complementary roles for grounded augmentation and paired consistency. Full details are in Table A6 and Appendix C.6. The controls address complementary explanations: update matching tests additional optimization, the same-data ablation isolates consistency, and leakage filtering tests direct answer cues. Their agreement provides stronger mechanism-level evidence than any diagnostic alone.

6 Conclusion

Taken together, the results show that CIDA combines targeted conflict detection, domain-grounded rewriting, quality filtering, and paired consistency training into an effective domain-adaptation pipeline. Across four benchmarks and three model families, its gains concentrate on high-conflict questions while low-conflict performance remains stable. Matched controls separate additional exposure, grounded augmentation, and paired invariance, and the analysis clarifies that agreement transfers prediction quality rather than guaranteeing correctness. CIDA is therefore best viewed as an answer-preserving domain-adaptation method; multi-hop and temporal tasks require stronger preservation constraints.

AI USE STATEMENT

Generative AI tools were used to support literature search, polish the writing, check theoretical derivations and proofs, and assist with the design and implementation of the experimental code, including analysis of synthetic-data and controlled experiments. All AI-assisted outputs were reviewed by the authors, who verified the claims, proofs, experimental results, figures, code, and references and take full responsibility for the final paper.

ETHICS STATEMENT

We use public benchmarks and one private domain dataset, which is withheld under privacy and access restrictions; only aggregate results are reported. We study controlled question answering, not personal identities or deployed decisions. Appendix I discusses broader impacts and Appendix J lists dataset and software licenses.

REPRODUCIBILITY STATEMENT

The pipeline is specified in Section 3: conflict detection (Section 3.2), knowledge-conditioned counterfactual generation (Section 3.3), quality filtering (Section 3.4), and the composite training objective (Section 3.5). The appendix reports dataset statistics and splits (C.1), conflict-score distributions (C.2), the principal training settings (C.3), matched-control definitions (C.4), counterfactual-generation statistics (C.5), the independent answer-recovery protocol (C.6) and its results (Table A6), rewriting prompts (E.1), filtering thresholds and pass rates (F), and the execution and compute-accounting scope (H). The motivating propositions are derived in Appendix G, with their scope specified in Appendix G.1. Three of the four datasets are public; CreditQA is proprietary and cannot be released, so the appendix documents its statistics and processing in enough detail to rebuild the pipeline on a comparable corpus.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=9-LSfSU74n-.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, and Antoine Bosselut. Tracking the limits of knowledge propagation: How LLMs fail at multi-step reasoning with conflicting knowledge. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5813–5847, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.273. URL https://aclanthology.org/2026.eacl-long.273/.

Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15540–15558. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghosal24a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11548–11559, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.733. URL https://aclanthology.org/2023.findings-acl.733/.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023. URL http://jmlr.org/papers/v24/23-0037.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Divyansh Kaushik, Eduard Hovy, and Zachary 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.

Mehran Kazemi, Sid Mittal, and Deepak Ramachandran. Understanding finetuning for factual knowledge extraction from language models, 2023. URL https://arxiv.org/abs/2301.11293.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25913–25935. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lai24c.html.

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 Hugo Larochelle,

Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Yougang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. KnowTuning: Knowledge-aware fine-tuning for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14535–14556, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.805. URL https://aclanthology.org/2024.emnlp-main.805/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

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, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559/.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. LLMs for generating and evaluating counterfactuals: A comprehensive study. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14809–14824, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.870. URL https://aclanthology.org/2024.findings-emnlp.870/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Jongyoon Song, Nohil Park, Bongkyu Hwang, Jaewoong Yun, Seongho Joe, Youngjune Gwon, and Sungroh Yoon. Entity-level factual adaptiveness of fine-tuning based abstractive summarization models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 915–929, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.55. URL https://aclanthology.org/2024.eacl-long.55/.

Sebastian Steindl, Ulrich Schäfer, and Bernd Ludwig. Counterfactual dialog mixing as data augmentation for task-oriented dialog systems. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4078–4087, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.363/.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ptvV5HGTNN.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Qihuang Zhong, Liang Ding, Xiantao Cai, Juhua Liu, Bo Du, and Dacheng Tao. KaFT: Knowledge-aware fine-tuning for boosting LLMs' domain-specific question-answering performance. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24085–24100, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1235. URL https://aclanthology.org/2025.findings-acl.1235/.

Symbol	Description	Section
$\mathcal{M}_\theta$	Pre-trained LLM with parameters θ	§3.1
$\mathcal{D}_{\text{sft}}$	Domain-specific SFT dataset	§3.1
(x_i, y_i)	Input question and target answer	§3.1
s_i	Conflict score for sample i	§3.1, Eq. 1
$\hat{y}_0$	Greedy decoding output (temperature 0)	§3.2
$\hat{y}_1, \ldots, \hat{y}_K$	Diverse sampling outputs	§3.2
τ_c	Consistency threshold	§3.2
$w_{\text{orig},i}$	Weight for original sample: $1 - \alpha_p \cdot s_i$	§3.5
$w_{\text{cf},i}$	Weight for counterfactual sample: $1 + \alpha_r \cdot s_i$	§3.5
α_p, α_r	Conflict penalty / reward coefficients	§3.5
$\tilde{x}_i$	Counterfactual input for sample i	§3.3
$\tilde{s}_i$	Score from reapplying the detector to a rewrite (diagnostic)	§3.4
τ_{cf}	Counterfactual generation threshold	§3.3
$\mathcal{L}_{\text{orig}}, \mathcal{L}_{\text{cf}}$	Weighted CE losses on original / counterfactual samples	§3.5
$\mathcal{L}_{\text{inv}}$	Counterfactual invariance loss (KL divergence)	§3.5, Eq. 4
λ_{inv}	Invariance loss weight	§3.5, Eq. 5
$\mathcal{L}_{\text{total}}$	Total training loss	Eq. 5
Q	Distribution of input pairs with a shared correct answer	§4
$\mathcal{G}_{\text{inv}}(\theta)$	Invariance gap	§4
R_x, R_{cf}	Expected bounded loss on original / rewritten inputs	Prop. 1
R_H	Smooth surrogate risk on a fixed high-conflict subset	Prop. 2

Table A1: Summary of notation used in Methods (§3) and Theory (§4).

A Notation Table

Table A1 summarizes the notation used throughout the paper.

B Extended Related Work

B.1 Knowledge Conflicts in Language Models

Knowledge conflicts arise when a language model's parametric knowledge contradicts external evidence or domain-specific facts. Longpre et al. (2021) first systematically studied this phenomenon, showing that pre-trained models often fail to update their internal knowledge when presented with conflicting evidence. Xu et al. (2024) categorized knowledge conflicts into three types: *context-memory conflicts* (between parametric and retrieved knowledge), *inter-context conflicts* (between multiple evidences), and *intra-memory conflicts* (within the model's parameters), and provided a comprehensive survey analyzing their causes, model behaviors under conflicts, and available solutions. Our work focuses on the first type—conflicts between pre-trained parametric knowledge and domain-specific facts in fine-tuning—and proposes a principled counterfactual augmentation approach to mitigate them.

Wang et al. (2024) proposed leveraging retrieval-augmented methods to resolve knowledge conflicts, but their approach operates at inference time without modifying the model's internal representations. Feng et al. (2026) showed that providing updated facts to LLMs can paradoxically worsen reasoning performance, revealing that models fail to faithfully integrate conflicting knowledge during multi-step inference. Our conflict detection module explicitly measures these biases through multi-temperature sampling consistency analysis. Mallen et al. (2023) demonstrated that LLMs struggle particularly with long-tail knowledge, where parametric memories are weakest—a finding that motivates our focus on detecting and mitigating conflicts for less well-known facts.

B.2 Understanding Fine-tuning and Factual Knowledge

A growing body of work investigates how fine-tuning affects the factual knowledge stored in LLMs. Kazemi et al. (2023) showed that fine-tuning primarily familiarizes the model with the QA format rather than injecting new knowledge. Ghosal et al. (2024) further identified a "frequency shock" phenomenon: fine-tuning on well-known facts outperforms fine-tuning on less-known facts, even when the goal is to improve performance on the latter. This finding aligns with our observation that

high-conflict samples (which often correspond to less well-known domain facts) require specialized treatment rather than simple exposure.

B.3 Data Selection and Sample Weighting for Fine-tuning

A line of work addresses training data quality issues through data selection and sample reweighting. Zhong et al. (2025) introduced Knowledge-aware Fine-tuning (KaFT), which assigns differential weights to samples based on a learned knowledge consistency score, with four weight tiers corresponding to quartile bins of the conflict distribution. Our approach differs from KaFT in two key aspects: (1) our conflict detection is conflict-annotation-free (it requires no labeled conflict signals, using only the task labels already present in the SFT data), and (2) we augment selected questions with grounded rewrites and jointly train their predictions.

Data-selection and filtering approaches more broadly treat conflicts as noise to be removed, whereas we view conflicts as opportunities for targeted intervention—by augmenting with counterfactuals, we teach the model to be robust to conflict rather than simply avoiding it.

FT-Top selects only the top-$k\%$ low-conflict samples for fine-tuning, based on the observation that models fine-tuned on high-quality subsets can outperform those trained on the full dataset. However, this approach sacrifices domain coverage and does not reduce the model's exposure to the conflicts it does encounter. KnowTuning (Lyu et al., 2024) identifies hard samples via perplexity and applies enhanced learning through two components: fine-grained knowledge augmentation and coarse-grained knowledge comparison. Its fine-grained knowledge augmentation requires an external reference model and a curated domain knowledge base to decompose answers into atomic facts and supply missing ones; neither is available in our setting—the proprietary CreditQA domain has no public reference model or released knowledge base, and we deliberately keep the pipeline free of such external dependencies so that it applies uniformly across all four benchmarks. We therefore compare against KnowTuning-Lite, which retains the perplexity-based hard-sample enhancement (the component reproducible under our constraints) but omits the fine-grained augmentation stage. We use the name KnowTuning-Lite throughout to make explicit that it is a reduced form and that our results do not claim superiority over the full KnowTuning method. Unlike either variant, CIDA constructs paired training inputs and adds an explicit prediction-consistency objective.

B.4 Knowledge Editing

Knowledge editing methods directly modify model parameters to update specific facts without full retraining. ROME (Meng et al., 2022) identifies and edits the MLP layers responsible for storing factual associations, treating knowledge as key-value pairs in the feed-forward networks. MEMIT (Meng et al., 2023) extends this to mass editing, updating thousands of facts simultaneously by distributing edits across multiple layers.

While these methods are effective for *atomic* factual updates (e.g., changing "The president of France is X" to "The president of France is Y"), they face fundamental limitations for domain-specific fine-tuning:

1. **Scalability**: Domain adaptation often requires updating thousands of interconnected facts, not isolated atomic edits. The cumulative effect of many edits can destabilize the model (Hoelscher-Obermaier et al., 2023).
2. **Relational knowledge**: Domain knowledge often involves complex relationships (e.g., "If LTV exceeds 70%, then additional collateral is required"), which cannot be captured by simple key-value edits.
3. **Unintended side effects**: Editing specific facts often causes degradation on related but unedited knowledge (Cohen et al., 2024), a phenomenon known as "edit ripple effects."

CIDA addresses conflicts at the *training data* level through grounded augmentation and paired training, with low-conflict accuracy remaining stable in the reported evaluations. Our approach is complementary to knowledge editing: while editing is suitable for post-hoc corrections of individual facts, CIDA is designed for systematic domain adaptation where conflicts are pervasive.

B.5 Counterfactual Data Augmentation

Counterfactual data augmentation has been explored for improving model robustness and reducing spurious correlations. Kaushik et al. (2020) first proposed learning the difference between original and counterfactual examples to reduce model sensitivity to spurious features.

Our approach differs from generic counterfactual augmentation in that we generate counterfactuals specifically *conditioned on retrieved domain knowledge*, guiding the rewritten inputs toward domain terminology and facts. Recent studies of LLM-generated counterfactuals (Nguyen et al., 2024) and counterfactual dialog mixing (Steindl et al., 2024) show that high-quality counterfactual generation is itself a nontrivial task, which motivates our retrieval-conditioned design. Closest to our setting, Neeman et al. (2023) use counterfactual data augmentation to train a QA model that disentangles parametric and contextual knowledge and emits both a contextual and a parametric answer at inference, and Song et al. (2024) apply controllable counterfactual augmentation to improve the entity-level factual adaptiveness of summarization models. Both operate on the answer/output side—one at inference-time knowledge separation, the other on entity-level factual signals for summarization—whereas CIDA selects samples by conflict score and injects, at training time, an invariance regularizer over paired original–counterfactual *inputs*, to reduce sensitivity to the surface cues varied by the rewrite.

B.6 Causal Inference and Invariant Learning

The design of CIDA draws from the invariant learning literature. Arjovsky et al. (2020) proposed Invariant Risk Minimization (IRM), which learns representations that are invariant across environments. Krueger et al. (2021) introduced Risk Extrapolation (REx), which reduces variance in risk across environments.

Our invariance loss $\mathcal{L}_{\text{inv}}$ is conceptually related to IRM, but instead of requiring multiple training environments, we construct "environments" via counterfactual intervention—the original sample and its counterfactual constitute two views of the same underlying factual content. This allows us to enforce invariance without collecting data from multiple domains.

Pearl (2009) formalized interventions in causal models. This perspective motivates varying input cues while preserving answer-relevant content. Following the counterfactual augmentation perspective of Kaushik et al. (2020), CIDA implements this idea through generated input pairs. The empirical analysis evaluates the effects of that training intervention on accuracy and prediction consistency.

B.7 Retrieval-Augmented Fine-tuning

RAG (Lewis et al., 2020) retrieves relevant documents at inference time and appends them to the input. While effective for knowledge-intensive tasks, RAG does not internalize the knowledge into model weights, creating a dependency on the retrieval system. Retrieval-augmented training can instead train models to attend to retrieved evidence (Izacard et al., 2023).

Our knowledge retrieval module serves a different purpose: rather than augmenting the input for inference, we use retrieved evidence as domain context for counterfactual generation. The evidence guides the LLM to rewrite questions in a way that redirects model attention from pre-trained priors to domain facts, producing training signal that reduces the model's conflict susceptibility rather than simply informing it at inference.

B.8 Regularization for Fine-tuning

Regularization techniques for fine-tuning aim to prevent catastrophic forgetting and maintain pre-trained capabilities. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) penalizes changes to important parameters determined by Fisher information. LoRA (Hu et al., 2022) constrains updates to low-rank adapters.

Our invariance regularizer is orthogonal to these weight-space regularizations: while EWC and LoRA constrain *how* the model changes, our invariance loss constrains *what* the model learns—specifically, it enforces that the model's predictions should be invariant under counterfactual input

Dataset	Train (original)	Train (sampled)	Val	Conflict Type
MedMCQA	182,822	15,487 / 18,383‡	4,183	Medical knowledge
PopQA	14,267	11,413 / 18,130‡	2,854	Long-tail factual
SciQ	12,679	11,679 / 15,183‡	1,000	Science domain
CreditQA†	–	6,937 / 9,662‡	237	Domain knowledge

†Domain-Specific

CreditQA cannot be released due to proprietary constraints. ‡Training subset size / total training samples after counterfactual augmentation.

Table A2: Dataset statistics.

Dataset	Low (<0.3)	Mid (0.3–0.7)	High (≥0.7)	Mean	Median
CreditQA	3.0%	65.0%	32.0%	0.614	0.623
MedMCQA	33.0%	50.0%	17.0%	0.436	0.374
PopQA	2.0%	45.0%	53.0%	0.696	0.702
SciQ	14.5%	55.5%	30.0%	0.512	0.498

Table A3: Unified conflict-score summary of the original training pools. Each row partitions its training pool into low, mid, and high conflict and sums to 100%. Mean and median refer to sample scores. High-conflict counts in Table A5 use this same partition, with percentages rounded for display.

perturbations that preserve the semantic content. This is a function-space regularization rather than a parameter-space one.

C Detailed Experimental Setup

C.1 Dataset Statistics

Table A2 provides detailed statistics for all datasets used in our experiments.

Sampling procedure for MedMCQA. MedMCQA contains $182{,}822$ training questions. We uniformly subsample within answer-topic strata to obtain the $15{,}487$-question training subset and then run conflict detection. Approximately 17% of this subset, or $2{,}633$ questions, belong to the high-conflict reporting bin $s_i \geq 0.7$ (Tables A3 and A5). The generation threshold $\tau_{\text{cf}} = 0.5$ also admits the upper part of the mid-conflict bin. Reporting-bin membership and generation eligibility therefore describe different quantities. PopQA, SciQ, and CreditQA use the same detection and augmentation procedure on their training pools without this additional subsampling.

C.2 Conflict Score Distribution

Table A3 summarizes the original training questions using one common partition: low ($0 \leq s_i < 0.3$), mid ($0.3 \leq s_i < 0.7$), and high ($0.7 \leq s_i \leq 1$). These boundaries are also used to name evaluation strata in Table 3, although the proportions below refer specifically to the training split. Means and medians summarize the scores themselves, not the bin centers. Table A5 uses the same high-conflict definition for its original-question counts.

C.3 Principal Hyperparameter Settings

C.4 Matched Controls and Ablation Interpretation

The controls in Table 2 distinguish training exposure, construction of additional examples, and coupling between original and rewritten questions. The final Vanilla SFT row is the original-budget reference from Table 1. The remaining rows use CIDA's training sample count and optimizer-step budget, the same seeds, and shared optimization hyperparameters. Matching the sample count concerns examples presented during training, including repeated originals in controls without new rewrites, rather than requiring the same number of distinct questions.

Exposure and construction controls. Update-matched Vanilla SFT uses only original question–answer examples at the matched training budget. Generic Paraphrase SFT uses answer-preserving

Module	Parameter	Setting
Conflict Detection	Number of decoded outputs (K)	3
	Sampling temperatures	$\{0, 0.7, 0.9\}$
	Internal-consistency threshold (τ_c)	0.7
Knowledge Retrieval	Sentence encoder	all-MiniLM-L6-v2
	Retrieved passages	Top-3
Counterfactual Gen.	Generation threshold (τ_{cf})	0.5
	Decoding	Low-temperature sampling
Quality Filter	Semantic threshold (τ_{sem})	0.7
	Knowledge-alignment threshold	0.5
	Bounded-change interval	$[0.05, 0.5]$
Training	Adaptation	LoRA on linear projections
	$\alpha_p, \alpha_r, \lambda_{\mathrm{inv}}$	0.7, 0.5, 0.5

Table A4: Principal settings used in the reported experiments. Standard optimizer and batching choices were shared across matched comparisons.

Dataset	Original	High-conflict originals	CF Kept	Augmented total
MedMCQA	15,487	2,633 (17.0%)	2,896	18,383
PopQA	11,413	6,049 (53.0%)	6,717	18,130
SciQ	11,679	3,504 (30.0%)	3,504	15,183
CreditQA	6,937	2,220 (32.0%)	2,725	9,662

Table A5: Original and augmented training-set sizes. "High-conflict originals" counts original questions with $s_i \geq 0.7$, using the same definition as Table A3. Generation eligibility is $s_i \geq \tau_{\mathrm{cf}} = 0.5$, and each selected question can produce 1–3 candidates. "CF Kept" counts retained rewrites across eligible questions, and "Augmented total" equals Original + CF Kept. Retained rewrite counts and high-conflict original counts have different denominators.

paraphrases and ordinary supervised fine-tuning, providing an alternative augmentation pipeline. Its comparison with Full CIDA evaluates the combined effect of construction, weighting, and paired training. The narrower *w/o Retrieval Grounding* comparison removes retrieved evidence from the rewriting condition while retaining the remaining CIDA components, directly examining grounding within the method. The *w/o CF Generation* row removes the rewritten examples and their associated supervised and invariance terms, measuring the end-to-end effect of augmentation.

Same-data invariance ablation. The *w/o Invariance Loss* row retains the original examples, grounded rewrites, target answers, and conflict weights of Full CIDA, with $\lambda_{\mathrm{inv}} = 0$. Both supervised terms remain active. Comparing this row with Full CIDA therefore measures the incremental effect of the regularizer on the same answer-supervised data. Its 2.55-point average difference is separate from the benefit of acquiring additional rewritten examples.

Pairing and component controls. Invariance on Random Pairs changes which rewritten question is assigned to an original in the consistency term while preserving the supervised question–answer examples. This breaks the semantic correspondence that motivates agreement. It is a negative control for indiscriminate regularization, whereas the same-data zero-invariance comparison establishes the incremental value of correctly paired training. The detection and filtering ablations remove the respective pipeline components under the matched budget. Their effects are conditional on the remaining components and are not added together as an attribution of the full improvement.

C.5 Counterfactual Generation Statistics

Table A5 shows the counterfactual generation and filtering statistics.

C.6 Independent Answer-Recovery Audit

Task and input isolation. The probe g_ϕ is trained independently of the CIDA model. Its task is to recover the reference answer using only the question and the context retained in the rewritten input.

Metric	Result
Answer n-gram occurrence in counterfactuals	$1.37\% \pm 0.42\%$
Independent answer-recovery probe AUC	0.53 ± 0.04
Answer-recovery gain: sanitized $\tilde{x}$ versus x	0.71 ± 0.52 pp
Change in CIDA main result after leakage filtering	$+0.24 \pm 0.29$ pp

Table A6: Answer-leakage audit (mean $\pm$ SD over three seeds). Probe inputs exclude answer fields and explicit answer markers. AUC measures answer-candidate discrimination, and recovery gain compares rewritten with original inputs. pp denotes percentage points.

Before probing, answer fields, label metadata, and explicit markers such as "the correct answer is" are removed. A gold answer is used as a training target or evaluation reference, never supplied through an answer field at test time. The same input-cleaning rule applies to original questions, which provide the reference condition for measuring the additional recoverability introduced by rewriting.

Data partition and prediction targets. The counterfactual pool is partitioned into 70% training, 15% validation, and 15% test data, stratified by dataset and conflict bin. Source questions are disjoint across splits, keeping an original and its rewritten views in the same partition. The training split fits the probe, the validation split supports model selection, and only held-out test predictions enter the reported results. The procedure is repeated with three random seeds. Multiple-choice prediction selects among the k supplied options. Open-ended prediction recovers an answer string from the sanitized input, as in PopQA. This separates independent answer recovery from the behavior of the CIDA-trained model.

Metrics and reference conditions. The AUC in Table A6 evaluates scores for correct versus incorrect answer candidates, with 0.5 denoting chance discrimination. It is distinct from multiple-choice accuracy, whose uniform-choice reference is $1/k$. A majority-choice predictor and predictors based on length or word-frequency features provide task-specific reference levels for interpreting recoverability. Original questions can themselves be answerable, so the primary comparison is the recovery increment from x_i to $\tilde{x}_i$, not an expectation that a capable reader must answer ordinary questions at chance. The reported 0.71-point difference summarizes this increment. Direct answer n-gram occurrence measures literal overlap in the rewritten question and added context, separately from the removed answer field and the supplied multiple-choice candidate list. The downstream filtering comparison then tests whether removing suspected overlap materially changes CIDA's result.

D Scope and Limitations, and Out-of-Scope Negative Results

CIDA applies consistency to rewrites intended to preserve the correct answer. Multi-hop and temporal questions make this requirement more demanding because intermediate entities and time qualifiers can determine the answer. We evaluate the present pipeline on two such benchmarks to characterize this boundary.

Multi-hop reasoning. HotpotQA requires combining facts across a reasoning chain. Rephrasing can preserve the answer, but changing an intermediate entity or relation can legitimately change it. Such changes must therefore be distinguished from surface-form perturbations when constructing invariant pairs.

Temporal knowledge. TimeQA asks about facts at specified times. A valid answer-preserving rewrite must retain the relevant time qualifier. Removing or changing that qualifier can invalidate a consistency target even when the resulting sentence remains fluent.

D.1 Results on HotpotQA and TimeQA

Table A7 compares vanilla fine-tuning and CIDA on HotpotQA (Yang et al., 2018) and TimeQA (Chen et al., 2021). These experiments assess the complete rewriting-and-training pipeline on multi-hop and temporal questions. The other baselines were not evaluated on these additional benchmarks.

Method	HotpotQA	TimeQA
Vanilla FT	87.58 ± 0.62	39.57 ± 1.15
CIDA	72.32 ± 0.84	33.48 ± 1.27
Δ vs. Vanilla FT	-15.26	-6.09

Table A7: Scope evaluation of the present CIDA pipeline on multi-hop and temporal QA. Entries are accuracy (%, mean $\pm$ SD over three seeds). The final row reports CIDA minus Vanilla FT in percentage points.

Accuracy decreases from 87.58% to 72.32% on HotpotQA and from 39.57% to 33.48% on TimeQA, changes of -15.26 and -6.09 points. These results establish a performance boundary for the present pipeline. A plausible explanation is that generated pairs sometimes alter answer-determining entities or time qualifiers, making consistency counterproductive. The experiments measure the combined effect of generation and training, rather than attributing the decrease to the regularizer alone. They motivate preserving reasoning-chain and temporal information explicitly when extending CIDA to these settings.

E Counterfactual Generation Details

E.1 Rewriting Prompts

We provide the shared rewriting template and the instruction substitutions for each conflict type.

Knowledge Conflict Prompt:

> You are a professional question rewriting assistant. Your task is to perform counterfactual rewriting of the question to reduce potential erroneous predictions by the model.
>
> This is a knowledge conflict question, where the model's prior knowledge differs from domain expertise. Please rewrite the question to: (1) Explicitly reference domain-specific terminology; (2) Eliminate expressions that may lead the model to rely on incorrect priors; (3) Make the intended answer more unambiguous.
>
> **Constraints:** (a) Preserve the core semantics; (b) The answer should remain consistent or more precise; (c) Do not directly reveal the answer; (d) The rewritten question should be natural and fluent.
>
> **Reference Knowledge:** {knowledge_condition}
>
> **Original Question:** {question}
>
> **Original Answer:** {answer}
>
> Output format: Rewritten Question: [rewritten question] Rewritten Answer: [consistent or refined answer]

Semantic Conflict Prompt: Similar structure, with instructions to: (1) Eliminate ambiguous expressions; (2) Make the question intent clearer; (3) Add necessary qualifiers.

Temporal Conflict Prompt: Similar structure, with instructions to: (1) Specify time range or version; (2) Emphasize the need for current/specific-period information; (3) Eliminate expressions that may trigger outdated knowledge.

E.2 Failure Cases

We analyze common failure modes of counterfactual generation:

- **Truncation** ($\sim$8% of generated counterfactuals): The LLM produces incomplete responses due to max_new_tokens limits. These are caught by Level 1 (readability) filtering.
- **Semantic drift** ($\sim$6%): The rewritten question changes the meaning beyond acceptable bounds. These are caught by Level 2 (semantic preservation) filtering.
- **Potential answer leakage**: Direct answer n-grams occur in $1.37\% \pm 0.42\%$ of counterfactuals. The independent answer-recovery audit in Appendix C.6 and Table A6 reports a small recovery increment over original inputs and a $+0.24$-point downstream change after leakage filtering.

Dataset	High Conflict	Mid Conflict	Overall
CreditQA	67.52%	82.34%	75.21%
MedMCQA	71.23%	85.12%	78.80%
PopQA	65.28%	80.76%	73.47%
SciQ	72.14%	86.38%	79.79%

Table A8: Quality filter pass rates by dataset and conflict level.

- **Insufficient conflict reduction** ($\sim$12%): Rewriting fails to sufficiently reduce the model-based conflict signal. The Level-4 semantic-change proxy screens candidate rewrites, while reapplying the detector provides the direct diagnostic described in Appendix F.

F Quality Filter Enhancement

F.1 Hybrid Filtering Strategy

The default Level-4 filter bounds semantic change. Its role is to screen rewrites efficiently; it does not itself recompute the conflict score. We explored a hybrid strategy that combines the fast proxy with the full conflict detector:

1. **Round 1 (Fast)**: Apply the semantic-based proxy filter to all counterfactuals;
2. **Round 2 (Precise)**: For counterfactuals passing Round 1 with original conflict score ≥ 0.7, apply the full conflict detector to verify actual conflict reduction.

This hybrid approach increases the precision of filtering from 72% (proxy only) to 91% (hybrid), at the cost of an additional $2\times$ forward passes for high-conflict samples ($\sim$35% of the data). We leave the production deployment of this hybrid strategy as future work.

F.2 Filter Pass Rates

G Derivations and Interpretation of the Motivating Analysis

G.1 Setting and Scope

The analysis uses a joint distribution $Q(X, \tilde{X}, Y)$ of paired inputs with a shared correct answer. The pair distribution defines the scope of the stability statement. A generated pair that changes the correct answer violates this interpretation even if its two predictions agree. For language generation, $P_\theta(\cdot \mid x)$ is the distribution over complete output answers under a common finite generation horizon, with an end-of-sequence convention. The task loss in Proposition 1 is bounded in $[0, 1]$. The smooth surrogate risk R_H in Proposition 2 is a separate object used only for the optimization analysis. No latent feature decomposition or inference from output agreement to an internal causal representation is needed.

G.2 Proof of Proposition 1

Proof. Fix $(x, \tilde{x}, y)$ and write $P = P_\theta(\cdot \mid x)$ and $\tilde{P} = P_\theta(\cdot \mid \tilde{x})$. Since $0 \leq \ell(a, y) \leq 1$, the variational characterization of total variation gives

$$|\mathbb{E}_{A\sim P}\ell(A, y) - \mathbb{E}_{A\sim\tilde{P}}\ell(A, y)| \leq \|P - \tilde{P}\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2}\mathrm{KL}(P\|\tilde{P})}. \tag{8}$$

The second inequality is Pinsker's inequality. Taking the expectation over Q, using the triangle inequality, and applying Jensen's inequality to the concave square root yields

$$\begin{aligned} |R_x(\theta) - R_{\mathrm{cf}}(\theta)| &\leq \mathbb{E}_Q\sqrt{\tfrac{1}{2}\mathrm{KL}(P_\theta(\cdot \mid X)\|P_\theta(\cdot \mid \tilde{X}))} \\ &\leq \sqrt{\mathcal{G}_{\mathrm{inv}}(\theta)/2}. \end{aligned} \tag{9}$$

This is precisely Eq. 6. In particular, $R_x(\theta) \leq R_{\mathrm{cf}}(\theta) + \sqrt{\mathcal{G}_{\mathrm{inv}}(\theta)/2}$. □

The bound transfers prediction quality within the specified pairs. Population generalization to a different distribution requires additional assumptions relating that distribution to Q. The bound is not applied directly to the unbounded cross-entropy objective or to the discontinuous accuracy of a greedy decoder.

G.3 Autoregressive Meaning of the KL Term

For an answer sequence $a_{1:T}$, autoregressive factorization gives $P_\theta(a_{1:T} \mid x) = \prod_{t=1}^{T} p_\theta(a_t \mid x, a_{<t})$. With padding after termination and a common horizon, the chain rule for KL divergence gives

$$\begin{aligned} &\mathrm{KL}(P_\theta(\cdot \mid x) \| P_\theta(\cdot \mid \tilde{x})) \\ &= \sum_{t=1}^{T} \mathbb{E}_{A_{<t} \sim P_\theta(\cdot|x)} \mathrm{KL}(p_\theta(\cdot \mid x, A_{<t}) \,\|\, p_\theta(\cdot \mid \tilde{x}, A_{<t}))\,. \end{aligned} \tag{10}$$

Each inner divergence compares full-vocabulary next-token distributions at the same answer prefix, rather than distributions at equal absolute prompt positions. This distinction matters because the original and rewritten prompts can have different lengths.

A teacher-forced surrogate instead evaluates the two next-token distributions at the shared gold-answer prefix $y_{<t}$ and averages over the T_y answer positions:

$$D_{\mathrm{TF}}(x, \tilde{x}; y) = \frac{1}{T_y} \sum_{t=1}^{T_y} \mathrm{KL}(p_\theta(\cdot \mid x, y_{<t}) \,\|\, p_\theta(\cdot \mid \tilde{x}, y_{<t}))\,. \tag{11}$$

Prompt and padding positions are outside this answer-token comparison. This expression describes a token-level surrogate, not the exact sequence divergence: it uses reference prefixes rather than the model-prefix expectation in the chain rule. Proposition 1 applies to the sequence divergence as defined, and does not identify a teacher-forced average with that quantity. In either formulation, cross-entropy provides the answer target and the divergence couples the predictions.

G.4 Proof of Proposition 2

Proof. For $d = \theta_p - \theta_b = -\eta\lambda_{\mathrm{inv}} g_{\mathrm{inv}}$, the descent lemma under the stated gradient-Lipschitz condition yields

$$R_H(\theta_b + d) \le R_H(\theta_b) + \langle \nabla R_H(\theta_b), d \rangle + \frac{L}{2}\|d\|^2. \tag{12}$$

Substitution gives

$$\begin{aligned} R_H(\theta_p) - R_H(\theta_b) \le &-\eta\lambda_{\mathrm{inv}} \langle \nabla R_H(\theta_b), g_{\mathrm{inv}} \rangle \\ &+ \frac{L\eta^2\lambda_{\mathrm{inv}}^2}{2}\|g_{\mathrm{inv}}\|^2, \end{aligned} \tag{13}$$

which is the stated inequality. The right-hand side is negative whenever $\langle \nabla R_H(\theta_b), g_{\mathrm{inv}} \rangle > (L\eta\lambda_{\mathrm{inv}}/2)\|g_{\mathrm{inv}}\|^2$. □

This is a sufficient local condition for the effect of an additional consistency update. The derivation keeps the curvature penalty and leaves gradient alignment explicit. It does not infer the alignment sign from semantic similarity, or extend a one-step comparison to uniform superiority of complete optimization trajectories. The matched ablations evaluate the corresponding empirical question under the training configuration used in this paper.

G.5 Interpretation of the Conflict Score

The score in Eq. 1 is an operational measure of stable target disagreement. Write $a(x) = 1 - \mathrm{Sim}(\hat{y}_0, y)$ and let $c_K(x)$ denote the measured sampling consistency. Then the score is exactly

$$s_K(x) = a(x)\left(\tfrac{1}{2} + \tfrac{1}{2}\mathbf{1}\{c_K(x) > \tau_c\}\right). \tag{14}$$

Its dependence on the threshold explains one source of sensitivity. If an estimator satisfies $c_K(x) \xrightarrow{p} c_*(x)$ and $c_*(x) \neq \tau_c$, then threshold decisions converge in probability, giving

$$s_K(x) \xrightarrow{p} a(x)\left(\tfrac{1}{2} + \tfrac{1}{2}\mathbf{1}\{c_*(x) > \tau_c\}\right). \tag{15}$$

To see this, set $\epsilon = |c_*(x) - \tau_c|/2$. Whenever $|c_K(x) - c_*(x)| < \epsilon$, the empirical and limiting consistency fall on the same side of the threshold. The probability of this event tends to one under the stated convergence condition. At the threshold, this argument does not apply.

The conditional limit preserves the actual piecewise score. It is not the model's probability of a wrong answer, and it does not establish finite-sample robustness for the default decoding configuration. Changes in temperatures, the sentence encoder, or the reference answer representation can change either factor. Fixed configurations make the reported within-dataset comparisons interpretable. The continuous training weights and the reporting bins use this same score.

H Compute Resources

Experiments were conducted on an internal multi-GPU accelerator node with sufficient memory for LoRA fine-tuning of 7–8B-parameter models. Multi-temperature detection and selective counterfactual generation were performed as offline preprocessing. The comparisons in Table 2 match sample exposure and optimizer steps, but not total preprocessing cost. Wall-clock time depends strongly on generation throughput and batching, so we do not report a single campaign-wide runtime.

I Ethics Statement and Broader Impacts

Data provenance. Three of the four evaluation datasets are standard public benchmarks: MedMCQA (Pal et al., 2022) (medical multiple-choice QA), PopQA (Mallen et al., 2023) (long-tail factual QA), and SciQ (Welbl et al., 2017) (science multiple-choice QA). None of them contain personally identifiable information (PII). The fourth dataset (CreditQA) is a proprietary domain-specific QA corpus; it has been internally reviewed and contains no PII, no customer-identifying fields, and no sensitive financial records. Usage of all public datasets follows their respective licenses (see Appendix J).

Positive impacts. CIDA targets a concrete safety-relevant failure mode of domain-specific LLM fine-tuning: hallucinations that arise when pre-trained priors conflict with domain facts. Reducing such hallucinations is particularly valuable in high-stakes domains such as healthcare, finance, and law, where incorrect-but-confident answers can cause direct harm. The consistency gains reported in Section 5.4 indicate that CIDA-trained models are less sensitive to surface-form rephrasing that preserves meaning, which is also a useful robustness property in deployment.

Potential negative impacts. Counterfactual generation is a dual-use capability. The same mechanism that rewrites questions to weaken spurious cues could, in principle, be repurposed to rewrite training data with the opposite intent, e.g., to inject biased priors, suppress a minority class, or obfuscate an undesirable behavior. We do not release a counterfactual-generation model; we only release training code that orchestrates standard LLM calls and retrieval.

Mitigations. Three built-in design choices mitigate misuse: (i) the four-stage quality filter (readability, semantic preservation, knowledge alignment, and bounded semantic change) rejects counterfactuals that drift from the original meaning or lose grounding; answer leakage is separately discouraged by the rewriting prompt and monitored in the failure analysis (Appendix E.2); (ii) the knowledge-conditioned rewriting step anchors each counterfactual to retrieved domain evidence, which makes it harder to inject fabricated content; and (iii) the training loss up-weights invariance rather than any particular answer, so the method is not a direct mechanism for forcing a specific output. We additionally recommend that deployments log the provenance of every counterfactual (source question, retrieved evidence, generator identity) to support post-hoc auditing.

Reproducibility and release. An anonymized code supplement accompanies the submission and contains the core conflict-detection, retrieval, counterfactual-generation, filtering, training-objective, and evaluation logic. The proprietary CreditQA dataset is not included; the three public benchmarks can be used with the provided data schema and pipeline. Full engineering wrappers and exact run manifests will be released after acceptance.

Asset	Type	License / Terms
MedMCQA (Pal et al., 2022)	Dataset	MIT
PopQA (Mallen et al., 2023)	Dataset	MIT
SciQ (Welbl et al., 2017)	Dataset	CC-BY-NC-3.0
CreditQA (ours)	Dataset	Proprietary; not released
HotpotQA (Yang et al., 2018)†	Dataset	CC-BY-SA-4.0
TimeQA (Chen et al., 2021)†	Dataset	BSD-3-Clause
Qwen3-8B (Yang et al., 2025)	Model	Apache-2.0
Llama-3-8B (Grattafiori et al., 2024)	Model	Llama 3 Community License
Mistral-7B (Jiang et al., 2023)	Model	Apache-2.0
HuggingFace Transformers	Software	Apache-2.0
FAISS	Software	MIT
sentence-transformers	Software	Apache-2.0
PyTorch	Software	BSD-3-Clause

Table A9: Licenses and usage terms for all external datasets, models, and software used in this paper. †HotpotQA and TimeQA are used only for the out-of-scope negative-result analysis in Appendix D, not for any CIDA result reported in the main text.

J Dataset and Software Licenses

Table A9 summarizes the license and usage terms for every external asset used in this paper. All usage is consistent with the respective license; CreditQA is the only asset that cannot be released.

K Limitations

CIDA is evaluated on four domain question-answering datasets and three model families. CreditQA is proprietary, so public reproduction must use the other three datasets or a comparable domain corpus. The current rewriting pipeline performs worse on HotpotQA and TimeQA (Appendix D), where preserving intermediate entities and temporal qualifiers is especially important. These results delimit the demonstrated scope of answer-preserving augmentation.

The conflict score is model-relative and depends on the sampling configuration and sentence encoder. The default configuration uses $K = 3$ outputs, and its finite-sample stability remains a limitation (Appendix G.5). Level 4 bounds semantic change, while the hybrid check in Appendix F applies the detector again. The generator receives the training answer, and the evidence for limited additional answer cues comes from the independent recovery and filtering audit in Table A6. This audit covers the specified probe and filtering conditions rather than every possible semantic shortcut.

The motivating analysis concerns bounded prediction loss on paired views and a local alignment condition. It does not convert the semantic proxy into a verified causal intervention or establish distribution-shift guarantees. The main evidence for effectiveness is the matched empirical comparison, including the same-data invariance ablation. Reported standard deviations summarize variation over three runs, and small differences in mean accuracy are interpreted descriptively.