Data Augmentation for Improving Contextual Faithfulness Under Knowledge Conflict

Mohith Rajesh* **Raj Maheshwari*** **Vashisth Tiwari**
Yonatan Bisk **Emma Strubell** **Jeremiah Milbauer** **Jared Fernandez**
Carnegie Mellon University
{mohithr, rajmahes}@andrew.cmu.edu

Abstract

External context enables language models to accomplish tasks which rely on information beyond what is captured during pretraining. A standard way to improve contextual faithfulness in these settings is to fine-tune on synthetic, modified questions for which the model must adhere to provided contexts inconsistent with information that may have been seen during training. While this approach reliably improves contextual faithfulness, we find that adapting to a narrowly targeted counterfactual signal yields degradation on other faithfulness evaluations and regression on tasks that evaluate a model's parametric world knowledge and reasoning. We find that data augmentation, in which a model is trained on *both* parametrically consistent and counterfactual examples improves performance on faithfulness tasks with reduced regression on parametric world knowledge and reasoning tasks. We find that these improvements are largely consistent across post-training objectives (i.e. SFT, DPO) and across ground-truth, counterfactual data augmentation ratios.

1 Introduction

Language models are frequently deployed in a changing world and experience performance regressions as parametric knowledge learned during pretraining becomes incorrect or outdated (Lazaridou et al., 2021; Luu et al., 2022; Dhingra et al., 2022). To address these challenges, language models may often incorporate external non-parametric evidence, e.g., retrieved passages in retrieval-augmented generation (Lewis et al., 2020). However, model performance can degrade when provided context is misleading or contradicts factual parametric knowledge encoded during pre-training (Mallen et al., 2023; Xu et al., 2024).

As models can over-rely on memorized knowledge when it conflicts with the provided context (Longpre et al., 2021), and LLMs show confirmation bias toward their parametric memory when evidence is mixed (Xie et al., 2024), recent work has sought to explicitly evaluate and train models to improve their *contextual faithfulness* (Bi et al., 2025; Wu et al., 2024; Ming et al., 2025; Si et al., 2026).

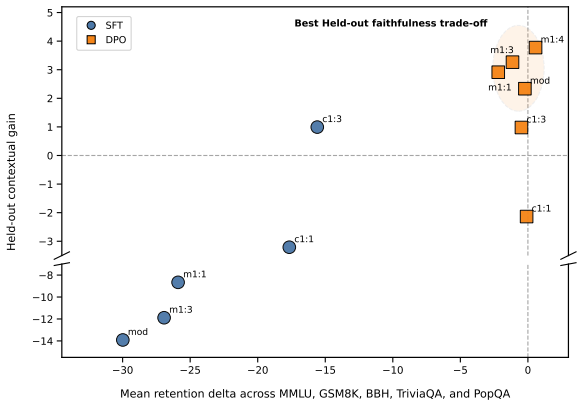


Figure 1: Data augmentation improves the trade-off between contextual faithfulness and parametric retention on ConFiQA. The y-axis shows aggregate held-out contextual gain (orig+mod where applicable) across knowledge-conflict benchmarks; the x-axis shows mean parametric retention on knowledge and reasoning benchmarks. Point labels: *mod* = modified-context examples only; *m* = multi-task; *c* = continual. Ratios (e.g., 1:1, 1:3, 1:4) denote the proportion of original- to modified-context examples in multi-task training; continual variants (c1:1, c1:3) follow a sequential schedule.

A common recipe to improve contextual faithfulness in LLMs is post-training on context-grounded QA data, often synthetic or counterfactual, via supervised fine-tuning, preference optimization, or reinforcement learning (Bi et al., 2025; Si et al., 2026; Li et al., 2025). These methods are typically trained and evaluated solely on questions that test the adherence to synthetic counterfactual context. This leaves two open concerns: do gains in contextual faithfulness come at the cost of forgetting the model's pre-trained parametric knowledge or gen-

*Equal contribution.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4419–4431
November 6-10, 2026 ©2026 Association for Computational Linguistics

eral reasoning behaviors, and do those gains transfer to held-out contextual benchmarks? We find that standard counterfactual-only training yields regressions during both transfer to tasks which evaluate different forms of faithfulness from the training demonstrations; and on tasks which evaluate a model's parametric ability to capture world knowledge or perform reasoning.

As an intervention to improve faithfulness and preserve parametric knowledge, we study *data augmentation* in which we mix original-context examples close to the model's pretraining and instruction-tuning distribution with the counterfactual examples used in faithfulness fine-tuning. We train **Llama-3.1-8B-Instruct** and **Qwen3-8B** on context-sensitive question-answer pairs from the ConFiQA dataset (Bi et al., 2025). We evaluate SFT and DPO-based post-training approaches across data augmentation ratios, measuring both contextual faithfulness and parametric retention. We consider data augmentation in two settings: (1) a multi-task learning setup in which the parametrically consistent examples are used alongside standard counterfactual-only training; and (2) a sequential, continual learning setup in which a model is first trained on parametrically consistent examples then on counterfactual examples.

Our contributions are: (i) a systematic comparison of data augmentation ratios for contextual-faithfulness fine-tuning; (ii) evidence that modified-only fine-tuning induces regression in a model's parametric knowledge and reasoning capabilities when evaluated in a context-free setting; and (iii) evidence that data augmentation improves transfer on held-out faithfulness benchmarks and incurs less parametric regression.

2 Related Work

Contextual-Faithfulness in Language Models. Retrieval of non-parametric information is frequently used to inform language models of information outside of their pretraining data (Lewis et al., 2020). However, retrieved non-parametric information can mislead models and conflict with their parametric knowledge, degrading downstream performance (Mallen et al., 2023).

As a result, a growing body of work evaluates how well LLMs ground answers in retrieved evidence when it conflicts with their parametric priors. This has yielded a number of datasets to evaluate the faithfulness in inconsistent, temporal, or misleading-evidence settings (Longpre et al., 2021; Xie et al., 2024; Bi et al., 2025; Wu et al., 2024; Ming et al., 2025; Ouyang et al., 2025; Su et al., 2024).

Training-time approaches include standard supervised finetuning on contextual faithfulness counter examples, direct preference optimization (DPO) on faithful versus stubborn responses (Bi et al., 2025), synthetic short-form QA with rules-based GRPO-based RL (Si et al., 2026), confidence-reasoning preference optimization (Huang et al., 2025), sentence-level generative and discriminative self-evolution (Li et al., 2025), and self-supervised context/no-context preferences (Tang et al., 2025). These approaches often consider a limited set of contextual-faithfulness types and neglect the impact of faithfulness training on parametric knowledge.

Data Augmentation and Mixtures. Variations in training data selection and structure are known to greatly affect the downstream task performance of language models. The efficacy of various transfer and data augmentation techniques is often dependent on the quality and relevance of the augmentations to the target task (Pruksachatkun et al., 2020; Vu et al., 2020). Relatedly, curriculum learning orders training examples from easier to harder (Bengio et al., 2009), and gradual domain adaptation adapts a model through a sequence of gradually shifting distributions rather than directly to the target (Kumar et al., 2020), motivating our continual-training variant in which a model is adapted first on parametrically consistent examples and then on counterfactual examples. Separately, data augmentation has long been used in NLP (Wei and Zou, 2019), and synthetic data is frequently used in modern large language model post-training to improve coverage and diversity of training data (Wang et al., 2023; Honovich et al., 2023). In contrast to synthetic data augmentation, we examine the effectiveness of augmenting training with ground-truth QA examples that are both parametrically consistent and contextually sensitive, whereas existing work on contextual faithfulness trains only on synthetic counterfactual instances. Additionally, the composition of training data strongly affects downstream behavior (Longpre et al., 2024), and data mixture and selection methods sample training data at different proportions or importance weights (Xie et al., 2023a,b).

3 Method

3.1 Problem Setup

To evaluate the effectiveness of post-training methods for contextual faithfulness, we consider the task of contextually sensitive question answering. Each example provides a question q, an original context-answer pair (c_o, a_o) that aligns with world knowledge, and a modified context-answer pair (c_m, a_m) in which the context has been counterfactually perturbed. We evaluate the QA performance of a model in its abilities to: correctly respond with the counterfactual answer when provided with modified context ($P_c(m)$); respond with the parametrically consistent answer when provided with the original context ($P_c(o)$) and when provided the question-only ($P_c(no_ctx)$).

All models are trained exclusively on the ConFiQA training split (Bi et al., 2025); as ConFiQA does not provide an official split, we construct our own train/test partition (Appendix C.3). We accordingly evaluate on: (i) *in-domain* contextual faithfulness on the ConFiQA test split; (ii) *held-out* contextual faithfulness on external conflict benchmarks constructed using different data pipelines (FaithEval, ClashEval, HoH, ConflictBank), testing whether models learn robust context-following behavior; and (iii) parametric retention in no-context settings on world-knowledge and reasoning tasks (TriviaQA, PopQA, MMLU, GSM8K, BBH). For ConFiQA we adopt the evaluation metrics of Bi et al. (2025), and for the held-out conflict benchmarks we use Partial Match; full details on datasets and evaluation metrics are provided in Appendix D.

3.2 Fine-Tuning for Contextual Faithfulness

We consider two training paradigms. **SFT** directly fine-tunes on supervised context-answer pairs with a language modeling objective. **DPO** follows the Context-DPO formulation (Bi et al., 2025), with the preferred response grounded in the provided context and the dispreferred response reflecting the competing parametric answer.

Our main intervention is *data augmentation*. Instead of training only on modified-context examples, we mix in original-context examples that are consistent with parametric information that may have been learned during pretraining. Original context examples are either mixed alongside the counterfactual examples to create a single larger training dataset in a multi-task configuration; or are used for sequential training in which the model is first adapted on original-context data and then continued on the modified-context objective.

We order the sequential intervention original-context first, then counterfactual, following a curriculum intuition: the model first adapts to the general format of contextually-sensitive QA on the prior-aligned, "easy" version of the task, then continues on the harder counterfactual case where it must follow context that conflicts with its priors. We place the counterfactual phase last so that training ends on the behavior we ultimately want, favoring context over parametric knowledge; the reverse order would end on prior-aligned data and partially undo the faithfulness adaptation. This mirrors curriculum- and gradual-adaptation approaches that order training from more in-distribution to more distribution-shifted data.

We compare modified-only training against variants with different original-context:modified-context ratios (e.g., 1:1, 1:3, and 1:4). In the multi-task configuration, the ratio is a mixture proportion within a single dataset: 1:3 includes three modified-context examples for every original-context example. In the continual configuration, the two example types occupy separate phases, and the ratio sets the number of epochs in each phase: 1:3 trains for one epoch on the full original-context set followed by three epochs on the full counterfactual set. Neither configuration is compute-matched with modified-only training: the multi-task variants add original-context examples on top of the full modified-context set, and the continual variants add original-context epochs before the counterfactual phase. We find, however, that performance (Figure 1) is driven by the augmentation ratio rather than the number of updates, and we return to this in the Limitations. Dataset construction details are in Appendix H.

4 Results and Discussion

In Tables 1, 2, and 3, we present the results of various post-training techniques for contextual faithfulness trained with different data augmentation ratios. Throughout, fine-tuned rows report the change relative to the base model, and cells are the mean over three seeds with the sample standard deviation. For ConFiQA, we define *contextual gain* as the average of $\Delta Pc(m)$ and $\Delta Pc(o)$, capturing performance under both modified (counterfactual) and original contexts, averaged across its QA, MC, and MR splits. For the held-out analysis,

Model	$Pc(m)\uparrow$	$Pc(o)\uparrow$	$Pc(\text{no_ctx})\uparrow$
Base	54.66 ±0.04	69.38 ±0.06	24.57 ±0.01
SFT (Modified only)	+34.43 ±0.04	+20.96 ±0.04	-3.34 ±0.05
SFT Multi-task (1:1)	+34.97 ±0.06	**+23.07** ±0.04	+8.22 ±0.12
SFT Multi-task (1:3)	+34.76 ±0.06	+22.17 ±0.16	+4.40 ±2.04
SFT Continual (1:3)	**+35.06** ±0.06	+22.32 ±0.05	**+9.61** ±0.08
SFT Continual (1:1)	+35.03 ±0.04	+22.30 ±0.07	+8.98 ±0.08
DPO (Modified only)	+19.51 ±0.06	+3.52 ±0.03	-7.87 ±0.13
DPO Multi-task (1:1)	+21.77 ±0.17	+6.23 ±0.19	-7.13 ±0.02
DPO Multi-task (1:3)	+20.21 ±0.03	+6.71 ±0.04	-13.04 ±0.13
DPO Multi-task (1:4)	**+22.93** ±0.07	**+9.71** ±0.04	-7.27 ±0.08
DPO Continual (1:1)	+17.73 ±0.06	+8.81 ±0.08	-2.70 ±0.13
DPO Continual (1:3)	+20.01 ±0.11	+9.53 ±0.17	**-1.98** ±0.11

Table 1: Data augmentation improves original-context faithfulness and parametric retention with little change to counterfactual faithfulness, most so for the continual variants. Results on ConFiQA, our in-domain counterfactual benchmark. We report contextual faithfulness under modified context ($Pc(m)$) and original context ($Pc(o)$), and no-context parametric retention ($Pc(\text{no_ctx})$).

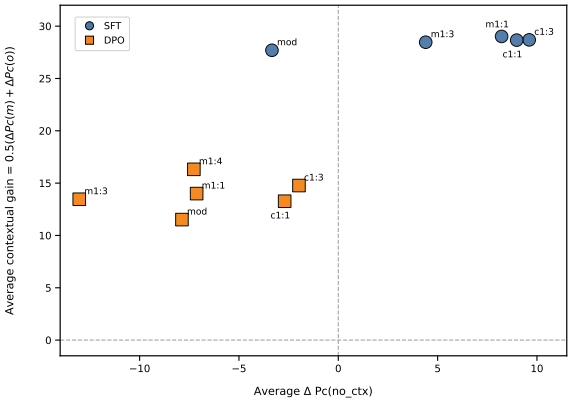


Figure 2: Aggregated in-domain Pareto comparison on ConFiQA. Each point averages performance across QA, MC, and MR split. Contextual gain is calculated by taking average of $\Delta Pc(m)$ and $\Delta Pc(o)$. Pc(no_ctx) is the question-only evaluation.

we analogously aggregate contextual-faithfulness gains across the non-ConFiQA knowledge conflict benchmarks, and use the no-context evaluations (TriviaQA, PopQA, MMLU, GSM8K, BBH) as probes of parametric retention.

4.1 Multi-task DPO Achieves Favorable Faithfulness–Retention Trade-offs

In-Domain (ConFiQA) Contextual Faithfulness and Retention. We consider the impact of data augmentation on ConFiQA test performance. In Figure 2, multi-task training improves contextual gain and reduces parametric knowledge regression relative to modified-only baselines. For DPO, multi-task training with ratios of 1:1 and 1:4 yields higher contextual gain with less severe retention loss. Notably, modified-only SFT also achieves

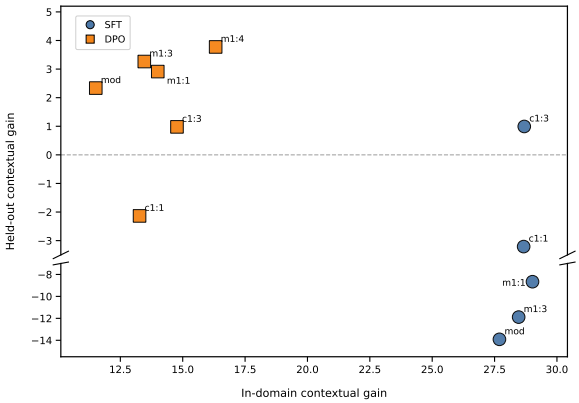


Figure 3: Relationship between in-domain and held-out contextual gains. The x-axis shows aggregate contextual gain on the ConFiQA test split and the y-axis shows aggregate contextual gain across held-out conflict benchmarks (FaithEval, ClashEval, HoH, ConflictBank).

strong in-domain contextual faithfulness (Table 1, $\Delta Pc(m)$ of $+34.43$), but as we show below these gains do not generalize to held-out benchmarks.

Generalization to Held-Out Benchmarks. In Figures 1 and 2, multi-task DPO improves contextual faithfulness on both the task it optimized for (ConFiQA) and held-out faithfulness benchmarks, while keeping parametric regression low. In contrast, modified-only DPO does not transfer as well to held-out benchmarks. Concretely, in Tables 2 and 3, DPO Multi-task (1:4) improves held-out faithfulness (e.g., FaithEval $+9.16$, ClashEval mod $+4.19$, HoH mod $+7.62$) while keeping parametric regression minimal (TriviaQA -0.48, BBH -0.17). Similarly, in Figure 3, multi-task DPO moves up and to the right relative to the modified-only baseline, improving both in-domain and held-out contextual faithfulness. This contrasts with modified-only SFT, which performs well in-domain but collapses on held-out benchmarks (Table 2; FaithEval -23.56, ClashEval orig -41.51).

Across Figures 1 and 3, the 1:4 setting attains the best overall trade-off, with the strongest aggregate held-out contextual gain and parametric regression close to zero, though the benefit of augmentation holds across all ratios. Continual-training DPO variants recover some retention but sit lower on the contextual-gain axis, placing them below the multi-task cluster on the overall frontier.

4.2 Continual Training Improves Parametric Retention for Reduced Contextual Gain

The continual-training variants occupy a distinct region of the faithfulness–retention frontier: they preserve, and in some cases improve, parametric

Model	FaithEval	ClashEval		HoH		ConflictBank	
	mod ↑	orig ↑	mod ↑	orig ↑	mod ↑	orig ↑	mod ↑
Base	54.07 ±0.16	88.18 ±0.03	78.71 ±0.02	73.87 ±0.13	74.68 ±0.02	74.57 ±0.39	70.08 ±0.51
SFT (Modified only)	-23.56 ±0.33	-41.51 ±0.17	-38.07 ±0.12	-9.08 ±0.87	-9.29 ±0.41	+12.44 ±0.66	+11.68 ±0.26
SFT Multi-task (1:1)	-17.06 ±0.18	-29.33 ±0.07	-23.50 ±0.13	-10.98 ±0.64	-11.13 ±0.45	+14.73 ±0.49	+16.64 ±0.14
SFT Multi-task (1:3)	-20.70 ±0.16	-35.45 ±0.15	-31.02 ±1.13	-13.15 ±0.26	-13.40 ±0.23	+15.56 ±0.89	+14.95 ±0.74
SFT Continual (1:3)	**+4.29** ±0.10	**-19.19** ±0.11	**-13.34** ±0.03	**-2.84** ±0.33	**-3.13** ±0.33	**+19.14** ±0.57	**+22.02** ±0.52
SFT Continual (1:1)	-2.38 ±0.24	-24.61 ±0.08	-19.47 ±0.09	-6.67 ±0.55	-6.91 ±0.41	+17.39 ±0.36	+20.19 ±0.46
DPO (Modified only)	+8.27 ±0.10	**-3.86** ±0.10	+2.59 ±0.03	+3.46 ±0.31	+3.58 ±0.03	+0.15 ±0.08	+2.16 ±0.15
DPO Multi-task (1:1)	**+11.03** ±0.18	-9.04 ±0.06	-0.57 ±0.21	**+8.03** ±0.32	**+8.12** ±0.26	**+0.16** ±0.07	**+2.66** ±0.08
DPO Multi-task (1:3)	+9.50 ±0.20	-6.56 ±0.10	+1.78 ±0.08	+7.86 ±0.11	+8.01 ±0.34	+0.14 ±0.09	+2.12 ±0.09
DPO Multi-task (1:4)	+9.16 ±0.24	-4.36 ±0.05	**+4.19** ±0.01	+7.58 ±0.18	+7.62 ±0.31	+0.14 ±0.07	+2.09 ±0.01
DPO Continual (1:1)	+6.33 ±0.20	-16.64 ±0.08	-10.27 ±0.19	+1.55 ±0.10	+2.03 ±0.22	+0.13 ±0.07	+1.92 ±0.03
DPO Continual (1:3)	+8.10 ±0.43	-10.82 ±0.08	-3.92 ±0.16	+5.35 ±0.08	+5.64 ±0.33	+0.16 ±0.08	+2.34 ±0.09

Table 2: Adding original-context examples improves transfer to held-out conflict benchmarks over modified-only training. Results on FaithEval, ClashEval, HoH, and ConflictBank, under original (orig) and modified (mod) context settings as applicable.

Model	MMLU ↑	GSM8K ↑	BBH ↑	TriviaQA ↑	PopQA ↑
Base	47.52 ±0.32	22.87 ±0.75	43.19 ±0.11	66.26 ±0.39	22.88 ±0.29
SFT (Modified only)	-36.52 ±0.46	-21.48 ±0.77	-41.16 ±0.13	-40.92 ±0.69	-9.85 ±0.38
SFT Multi-task (1:1)	-28.97 ±0.39	-19.79 ±0.72	-40.15 ±0.15	-36.29 ±0.68	-4.27 ±0.24
SFT Multi-task (1:3)	-32.37 ±1.72	-19.96 ±0.74	-41.72 ±0.13	-34.33 ±1.85	-6.27 ±1.20
SFT Continual (1:3)	**-11.42** ±0.57	-14.81 ±0.90	**-34.49** ±0.12	**-17.46** ±0.61	**+0.22** ±0.23
SFT Continual (1:1)	-15.15 ±0.85	**-13.27** ±0.95	-35.87 ±0.20	-21.42 ±0.53	-2.60 ±0.24
DPO (Modified only)	+0.70 ±0.18	+3.59 ±0.62	**+0.02** ±0.10	-2.67 ±0.42	-2.76 ±0.18
DPO Multi-task (1:1)	+0.10 ±0.25	+2.73 ±0.92	-9.03 ±0.13	-3.41 ±0.45	-1.37 ±0.15
DPO Multi-task (1:3)	**+1.25** ±0.23	+6.65 ±1.03	-1.18 ±0.19	-5.76 ±0.47	-6.66 ±0.08
DPO Multi-task (1:4)	-1.03 ±0.24	**+6.67** ±0.79	-0.17 ±0.11	**-0.48** ±0.30	-2.16 ±0.07
DPO Continual (1:1)	-0.81 ±0.03	+2.68 ±0.84	-0.05 ±0.19	-1.00 ±0.12	-1.30 ±0.09
DPO Continual (1:3)	-0.11 ±0.22	+0.33 ±0.69	-0.31 ±0.06	-1.13 ±0.32	**-1.17** ±0.08

Table 3: Data augmentation for contextual faithfulness minimally impacts parametric knowledge and reasoning. Closed-book MMLU, GSM8K, and BBH, and factual recall on TriviaQA and PopQA.

knowledge more effectively than their modified-only and multi-task counterparts, but at the cost of reduced contextual gains. In Figure 2, continual SFT is particularly striking in that it *both* improves faithfulness and parametric knowledge (Table 1, $\Delta Pc(m)$ of $+35.06$ alongside $\Delta Pc(\text{no_ctx})$ of $+9.61$). Among DPO variants, continual DPO achieves the best in-domain parametric retention (Table 1, $\Delta Pc(\text{no_ctx})$ of -1.98), for lower contextual gain. This pattern extends to held-out benchmarks: in Figure 1, continual SFT improves both retention and contextual gain more strongly than multi-task SFT, with Tables 2 and 3 showing near-zero parametric regression on PopQA ($+0.22$) alongside a positive FaithEval gain ($+4.29$) for the 1:3 ratio. Continual DPO, by contrast, retains comparably or better than other DPO variants but incurs substantially lower contextual gains than both multi-task and modified-only DPO.

4.3 Robustness Across Models

To test whether these findings are model-specific, we evaluate a representative subset of settings on Qwen3-8B (modified-only, one multi-task ratio, and one continual ratio per method; Appendix E, Table 4). The DPO results replicate Llama: multi-task DPO (1:4) improves contextual faithfulness over modified-only DPO across both in-domain and held-out benchmarks while reducing parametric regression on TriviaQA, PopQA, MMLU, and GSM8K. The in-domain SFT also observes consistent behaviors, with SFT variants reaching similar $Pc(m)$ gains. The main difference is on held-out SFT: Qwen3 modified-only SFT does not exhibit the sharp degradation seen on Llama, so augmentation is less necessary there, though data augmentations still offers performance improvements on ConflictBank and multi-task is the only SFT setting to improve factual recall on PopQA ($+3.90$).

Taken together, these results highlight that different training strategies interact differently with each fine-tuning objective. Rather than a single dominant strategy, the optimal choice depends on the objective itself: sequential continual training proves especially effective when the objective is aggressive, as in SFT, while multi-task better serves more conservative objectives like DPO, a pattern that largely holds across both models we study.

Limitations

In our work, we examine the effectiveness of parametrically consistent and counterfactual data mixtures on the contextual faithfulness of large language models. We consider this from the perspec-

tive of multi-task and sequential training. However, we acknowledge that there are other forms of training which could incorporate data mixtures.

Our augmented settings are not compute-matched with modified-only training, since they add original-context examples (multi-task) or additional epochs (continual) on top of the full counterfactual set. We find that position on the faithfulness–retention frontier is governed by the augmentation ratio rather than the number of examples or updates (Appendix H). Our experiments also are limited to investigation of 8B parameter instruction-tuned models. Our contribution is a training-time intervention that leaves inference untouched, and we do not investigate training-free approaches to contextual faithfulness, such as inference-time decoding methods (e.g., watermark-inspired or context-aware decoding). These approaches are orthogonal to ours as they operate on different stages of the model pipeline, and as a result we would expect the gains to stack. A systematic study of combining our data mixtures with training-free decoding methods is outside the scope of this work, though we view it as a promising direction.

Additionally, we acknowledge that improving contextual faithfulness and instruction-following capabilities of language models could expose safety risks, for instance through increased susceptibility to jailbreaking.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. In *The Thirteenth International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ananya Kumar, Tengyu Ma, and Percy Liang. 2020. Understanding self-training for gradual domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5468–5479. PMLR.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Kun Li, Tianhua Zhang, Yunxiang Li, Hongyin Luo, Abdalla Mohamed Salama Sayed Moustafa, Xixin Wu, James R. Glass, and Helen M. Meng. 2025. Generate, discriminate, evolve: Enhancing context faithfulness via fine-grained sentence-level self-evolution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17091–17105, Vienna, Austria. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *International Conference on Learning Representations*, pages 29430–29456.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. HoH: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6036–6063, Vienna, Austria. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Shuzheng Si, Haozhe Zhao, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Bofei Gao, Kangyang Luo, Wenhao Li, Yufei Huang, Gang Chen, Fanchao Qi, Minjia Zhang, Baobao Chang, and Maosong Sun. 2026. Teaching large language models to maintain contextual faithfulness via synthetic tasks and reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33001–33009.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268. Curran Associates, Inc.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them.

Xiaqiang Tang, Yi Wang, Keyu Hu, Rui Xu, Chuang Li, Weigao Sun, Jian Li, and Sihong Xie. 2025. SSFO: Self-supervised faithfulness optimization for retrieval-augmented generation.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Kevin Wu, Eric Wu, and James Zou. 2024. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *Advances in Neural Information Processing Systems*, volume 37, pages 33402–33422. Curran Associates, Inc.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023a. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818. Curran Associates, Inc.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023b. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

A Data and Model Usage & Licensing

The Llama-3.1 model is used under the terms of the Llama 3 Community License Agreement*, and Qwen3-8B is used under the Apache 2.0 License. The ClashEval, PopQA, MMLU, GSM8K, and BBH datasets are used under the MIT License. The FaithEval, TriviaQA, and HoH datasets are used under the Apache 2.0 License. ConflictBank is used under the CC-BY 4.0 License.

To our knowledge the data used for model training and evaluation does not contain PII or offensive content, as datasets are largely synthetically generated or evaluate for standard world knowledge. Models are trained on English language data. The use of data is consistent with the intended use.

B Benchmarks

We train on ConFiQA and evaluate on ConFiQA, four held-out knowledge-conflict benchmarks, and five closed-book benchmarks. Datasets with more than 10,000 examples are shuffled with a fixed seed and truncated to 10,000; smaller ones are used in full. Prompting and scoring are described in Appendix D.

Training and in-domain evaluation. ConFiQA (Bi et al., 2025) pairs each original question, context, and answer with a counterfactual version in which the answer is replaced by an entity of the same type and the context is edited to support it. It has three subsets of 6,000 examples each: single-hop QA, multi-hop reasoning (MR), where the substituted fact sits in a 2–4 hop reasoning chain, and multiple conflicts (MC), where several facts in the chain are substituted. We train on 4,500 examples per subset and test on the remaining 1,500 (Appendix C.3), reporting each subset separately.

Held-out knowledge-conflict benchmarks.

- **FaithEval** (Ming et al., 2025) tests faithfulness under unanswerable, inconsistent, and counterfactual contexts. We use the counterfactual task: about 1K multiple-choice science questions from ARC-Challenge, each paired with an LLM-written multi-paragraph context that argues for a wrong option (three to five options per question). There is no original context, so we report only the modified setting.

*https://www.llama.com/llama3/license/

Model	ConFiQA			FaithEval	ClashEval		HoH		ConflictBank		Closed-book / recall				
	$Pc(m)\uparrow$	$Pc(o)\uparrow$	$Pc(\text{no_ctx})\uparrow$	mod ↑	orig ↑	mod ↑	orig ↑	mod ↑	orig ↑	mod ↑	MMLU ↑	GSM8K ↑	BBH ↑	TriviaQA ↑	PopQA ↑
Base	60.87 ±0.05	70.21 ±0.01	20.95 ±0.11	47.74 ±0.21	87.30 ±0.02	76.00 ±0.08	61.95 ±0.56	62.39 ±0.29	74.67 ±0.48	69.05 ±0.46	54.81 ±0.25	16.78 ±0.19	42.73 ±0.13	46.74 ±0.23	14.34 ±0.35
SFT (Modified only)	+28.26 ±0.05	+18.48 ±0.01	+8.12 ±0.22	+19.07 ±0.21	**-4.62** ±0.05	-3.68 ±0.22	**+20.08** ±0.59	**+20.32** ±0.49	+0.27 ±0.08	+4.64 ±0.08	-3.48 ±0.05	+3.41 ±0.13	-27.81 ±0.15	**-1.05** ±0.30	-0.83 ±0.33
SFT Multi-task (1:1)	+28.64 ±0.05	**+22.24** ±0.01	+11.30 ±0.10	**+19.82** ±0.27	-10.01 ±0.13	**-3.43** ±0.15	+15.28 ±0.82	+15.24 ±0.47	+17.54 ±0.48	**+21.95** ±0.53	-5.04 ±0.10	+7.99 ±0.37	-22.48 ±0.08	-6.04 ±0.17	+1.77 ±0.25
SFT Multi-task (1:3)	+28.58 ±0.06	+21.41 ±0.01	**+15.66** ±0.16	+16.82 ±0.29	-10.06 ±0.09	-5.23 ±0.19	+11.82 ±0.39	+11.87 ±0.53	+16.55 ±0.66	+20.14 ±0.45	-3.83 ±0.19	**+10.46** ±0.23	**-20.99** ±0.10	-1.76 ±0.28	**+3.90** ±0.31
SFT Continual (1:3)	**+28.90** ±0.05	+21.00 ±0.02	+3.76 ±0.08	+17.98 ±0.62	-14.39 ±0.08	-11.57 ±0.11	+11.32 ±0.82	+12.18 ±0.44	**+18.51** ±0.73	+21.12 ±0.48	**-3.43** ±0.21	+8.26 ±0.27	-24.38 ±0.10	-9.82 ±0.07	-0.73 ±0.40
DPO (Modified only)	+8.92 ±0.11	-1.88 ±0.08	-6.27 ±0.08	+7.46 ±0.31	-0.72 ±0.03	+2.55 ±0.13	+0.57 ±0.43	+0.58 ±0.37	+0.13 ±0.04	+1.40 ±0.10	+0.02 ±0.06	+0.61 ±0.20	-0.68 ±0.08	-9.07 ±0.11	-3.57 ±0.11
DPO Multi-task (1:3)	**+12.48** ±0.03	+2.40 ±0.10	-4.90 ±0.05	**+8.68** ±0.31	-0.30 ±0.02	+4.31 ±0.10	**+3.31** ±0.33	**+3.41** ±0.29	+0.14 ±0.05	+1.58 ±0.14	-0.05 ±0.04	+1.01 ±0.29	+0.92 ±0.19	-7.21 ±0.22	-2.99 ±0.07
DPO Multi-task (1:4)	+12.13 ±0.07	+2.02 ±0.04	-4.74 ±0.17	+7.80 ±0.16	**+0.71** ±0.06	**+5.12** ±0.13	+2.77 ±0.48	+2.66 ±0.22	+0.14 ±0.05	+1.46 ±0.10	**+0.12** ±0.06	+0.94 ±0.29	-0.72 ±0.18	**-6.02** ±0.25	-2.90 ±0.07
DPO Continual (1:3)	+10.05 ±0.00	**+2.67** ±0.08	**-4.47** ±0.14	+7.66 ±0.27	-0.22 ±0.04	+3.43 ±0.08	+0.63 ±0.21	+0.69 ±0.23	**+0.16** ±0.06	**+1.74** ±0.15	+0.07 ±0.07	**+2.10** ±0.16	**+1.99** ±0.13	-6.34 ±0.14	**-2.77** ±0.07

Table 4: Qwen3-8B results across all benchmarks, laid out to match the Llama-3.1-8B-Instruct tables. Fine-tuned rows are deltas relative to the Qwen3-8B base model; cells are the mean over three seeds with the sample standard deviation as a subscript. Best values are bolded within each training family.

	ConFiQA						
Model	QA	MC	MR	ClashEval	HoH	TriviaQA	PopQA
Llama-3.1-8B-Instruct							
Base	4.0	39.9	41.5	32.2	49.1	1.5	16.3
SFT (all variants)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DPO (Modified only)	20.6	74.2	74.2	56.1	71.7	8.0	39.8
DPO Multi-task (1:1)	27.2	58.8	58.8	67.2	74.2	13.3	47.6
DPO Multi-task (1:3)	43.2	93.4	93.0	87.9	90.7	13.1	60.7
DPO Multi-task (1:4)	12.4	64.2	66.7	60.8	73.9	3.4	32.0
DPO Continual (1:1)	15.1	58.0	56.8	59.3	72.8	7.8	27.6
DPO Continual (1:3)	13.7	57.1	56.4	55.0	70.1	7.3	34.0
Qwen3-8B							
Base	15.9	48.1	48.6	47.6	63.5	11.0	56.2
SFT (Modified only)	0.2	0.7	0.5	5.4	14.5	0.2	0.6
SFT Multi-task (1:1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SFT Multi-task (1:3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DPO (Modified only)	32.6	60.8	61.9	58.5	73.8	16.8	67.3
DPO Multi-task (1:3)	32.7	59.1	58.7	56.3	71.6	17.5	67.2
DPO Multi-task (1:4)	32.0	59.0	60.1	55.2	70.9	16.2	66.3
DPO Continual (1:3)	32.0	60.7	62.1	63.0	76.7	20.0	68.3

Table 5: No-context abstention rate (%, `INSUFFICIENT_CONTEXT` responses), mean over three seeds.

- **ClashEval** (Wu et al., 2024) contains over 1,200 questions from six domains, such as drug dosages, Olympic records, and locations, each paired with a reference document. The answer inside the document is perturbed with edits ranging from subtle to blatant, which gives an original and a modified context per question.

- **HoH** (Ouyang et al., 2025) targets outdated information. Its QA pairs are built from real Wikipedia revisions with a token-level diff and an LLM pipeline, where each QA pair has the current evidence and answer with earlier, outdated versions. We use the current evidence as the original context and the first outdated version as the modified context, then score each against its own answer. Unlike the original HoH setup, which always expects the current answer, we instruct our model to follow whichever version it is shown.

- **ConflictBank** (Su et al., 2024) is built from Wikidata facts and covers conflicts caused by misinformation, temporal change, and semantic divergence. Conflicting claims are made by substituting the object of a fact, and an LLM writes supporting evidence for each claim. Questions are four-option multiple choice. We use the default evidence as the original context and the misinformation-conflict evidence as the modified context.

Closed-book benchmarks. These are evaluated without context and measure parametric retention.

- **MMLU** (Hendrycks et al., 2021), all 57 subjects): multiple-choice world knowledge.

- **GSM8K** (Cobbe et al., 2021) (`gsm8k`): grade-school math word problems, scored on the final numeric answer.

- **BBH** (Suzgun et al., 2022) (`lukaemon/bbh`, all 27 tasks): multi-step reasoning.

- **TriviaQA** (Joshi et al., 2017): factual recall. We use the validation split because test answers are not public.

- **PopQA** (Mallen et al., 2023): factual recall over entities of varying popularity, including long-tail facts.

C Training and Implementation Details

C.1 Exact Prompt Templates

Unless otherwise noted, all of our main runs use the same prompt schema.

The exact open-ended QA prompt with context is:

```
Answer the question using only the provided
↪  context.

Rules:
1. Use only facts explicitly stated in the
↪  context.
2. Do not use prior knowledge, world knowledge,
↪  or assumptions.
3. Base the answer on the single span in the
↪  context that most directly answers the
↪  question.
4. If the context does not explicitly contain the
↪  answer, return exactly: INSUFFICIENT_CONTEXT
```

Model	QA split			MC split			MR split		
	$Pc(m)\uparrow$	$Pc(o)\uparrow$	Pc(no_ctx) $\uparrow$	$Pc(m)\uparrow$	$Pc(o)\uparrow$	Pc(no_ctx) $\uparrow$	$Pc(m)\uparrow$	$Pc(o)\uparrow$	Pc(no_ctx) $\uparrow$
Base	72.74 $\pm$0.10	82.09 $\pm$0.08	43.91 $\pm$0.08	40.47 $\pm$0.12	62.69 $\pm$0.10	14.98 $\pm$0.14	50.78 $\pm$0.10	63.36 $\pm$0.21	14.82 $\pm$0.08
SFT (Modified only)	+25.59 $\pm$0.10	+13.16 $\pm$0.10	-15.07 $\pm$0.13	+44.07 $\pm$0.12	+23.60 $\pm$0.07	+0.62 $\pm$0.14	+33.62 $\pm$0.10	+26.11 $\pm$0.21	+4.42 $\pm$0.04
SFT Multi-task (1:1)	+25.15 $\pm$0.13	**+14.91** $\pm$0.10	-5.27 $\pm$0.07	+44.40 $\pm$0.12	**+26.64** $\pm$0.10	+13.96 $\pm$0.10	**+35.36** $\pm$0.10	**+27.64** $\pm$0.21	+15.98 $\pm$0.27
SFT Multi-task (1:3)	+25.35 $\pm$0.27	+13.82 $\pm$0.08	-3.80 $\pm$1.73	+44.11 $\pm$0.04	+25.47 $\pm$0.24	+7.98 $\pm$1.81	+34.82 $\pm$0.14	+27.22 $\pm$0.27	+9.02 $\pm$2.58
SFT Continual (1:3)	**+25.79** $\pm$0.10	+14.56 $\pm$0.10	**-0.73** $\pm$0.18	+44.47 $\pm$0.12	+25.51 $\pm$0.10	+13.40 $\pm$0.18	+34.91 $\pm$0.10	+26.89 $\pm$0.20	**+16.18** $\pm$0.23
SFT Continual (1:1)	+25.53 $\pm$0.10	+14.38 $\pm$0.08	-1.33 $\pm$0.13	**+44.53** $\pm$0.12	+25.49 $\pm$0.10	**+14.00** $\pm$0.31	+35.02 $\pm$0.10	+27.04 $\pm$0.21	+14.27 $\pm$0.00
DPO (Modified only)	+10.47 $\pm$0.00	+0.09 $\pm$0.14	-5.51 $\pm$0.34	+31.36 $\pm$0.14	+5.51 $\pm$0.04	-9.76 $\pm$0.21	+16.69 $\pm$0.04	+4.96 $\pm$0.14	-8.36 $\pm$0.23
DPO Multi-task (1:1)	**+15.19** $\pm$0.08	+2.44 $\pm$0.20	-3.78 $\pm$0.14	+31.53 $\pm$0.23	+8.60 $\pm$0.41	-9.40 $\pm$0.31	+18.60 $\pm$0.35	+7.64 $\pm$0.14	-8.22 $\pm$0.28
DPO Multi-task (1:3)	+14.28 $\pm$0.13	+3.67 $\pm$0.18	-13.64 $\pm$0.30	+28.96 $\pm$0.14	+8.22 $\pm$0.17	-13.11 $\pm$0.14	+17.40 $\pm$0.27	+8.24 $\pm$0.17	-12.36 $\pm$0.04
DPO Multi-task (1:4)	+13.28 $\pm$0.12	+4.02 $\pm$0.17	-3.93 $\pm$0.13	**+34.80** $\pm$0.13	**+12.78** $\pm$0.14	-9.09 $\pm$0.28	**+20.71** $\pm$0.04	**+12.33** $\pm$0.23	-8.80 $\pm$0.12
DPO Continual (1:1)	+10.34 $\pm$0.18	**+5.16** $\pm$0.08	-2.53 $\pm$0.27	+26.64 $\pm$0.17	+10.73 $\pm$0.07	-3.62 $\pm$0.20	+16.20 $\pm$0.13	+10.53 $\pm$0.29	**-1.93** $\pm$0.18
DPO Continual (1:3)	+11.16 $\pm$0.21	+4.47 $\pm$0.00	**-0.29** $\pm$0.08	+30.27 $\pm$0.12	+12.73 $\pm$0.18	**-3.24** $\pm$0.20	+18.60 $\pm$0.18	+11.38 $\pm$0.39	-2.40 $\pm$0.07

Table 6: Llama-3.1-8B-Instruct ConFiQA results by split. Base row is absolute; fine-tuned rows are deltas relative to the base model. Mean over three seeds with sample standard deviation as a subscript. Best values are bolded within each training family.

Method	Setting	Examples seen	#Updates
SFT	modified-only	13,495	211
	multi-task (1:1)	26,995	422
	multi-task (1:3)	17,995	282
	continual (c1:1)	80,985	1,266
	continual (c1:3)	53,985	844
DPO	modified-only	40,494	2,532
	multi-task (1:1)	80,994	5,064
	multi-task (1:3)	53,994	3,375
	multi-task (1:4)	50,619	3,165
	continual (c1:1)	80,994	5,064
	continual (c1:3)	54,094	3,382

Table 7: Per-setting training budget. Examples seen is the cumulative number of training examples across all epochs (dataset size $\times$ epochs); #Updates is the total number of optimizer steps. Augmented settings add original-context examples on top of the full modified-context set and are therefore not compute-matched with modified-only training.

```
5. If the context contains conflicting
↪  information, choose the statement that most
↪  directly answers the question and do not
↪  reconcile with outside knowledge.
6. Reply with ONLY the answer itself: no
↪  preamble, no explanation, and no repetition
↪  of the question.

Formatting:
- If the answer is numerical, return digits only.
- Do not use commas in numbers.
- Do not add units unless they appear in the
↪  answer span itself.

Context:
'''{context}'''

Question: {question}
```

The exact no-context open-ended prompt is:

```
Answer the question using only your own
↪  knowledge.

Rules:
1. Choose the single answer that most directly
↪  answers the question.
2. If you cannot answer confidently from your
↪  knowledge, return exactly:
↪  INSUFFICIENT_CONTEXT
3. If the question is ambiguous, choose the
↪  interpretation that is most directly
↪  supported by the wording of the question.
4. Reply with ONLY the answer itself: no
↪  preamble, no explanation, and no repetition
↪  of the question.

Formatting:
- If the answer is numerical, return digits only.
- Do not use commas in numbers.
- Do not add units unless they appear in the
↪  answer span itself.

Question: {question}
```

The exact multiple-choice prompt with context is:

```
You are given:
1. Context retrieved from a document.
2. A multiple-choice question with answer
↪  choices.

Your task is to select the answer choice that is
↪  most strongly supported by the provided
↪  context.
Rely only on the context, even if another choice
↪  may seem correct based on prior knowledge.
Do not use outside knowledge, common sense beyond
↪  the context, or assumptions.
If multiple choices seem plausible, choose the
↪  one with the closest and strongest grounding
↪  in the context.

Context:
{context}

Question:
{question}

Choices:
{choices}

Respond with ONLY the single letter of the
↪  correct choice ({final_answer_format}). No
↪  words, no punctuation, no explanation, no
↪  restating of the choice.
```

The exact no-context multiple-choice prompt is:

```
Answer the following multiple-choice question.

Question:
{question}

Choices:
{choices}

Respond with ONLY the single letter of the
↪ correct choice ({final_answer_format}). No
↪ words, no punctuation, no explanation, no
↪ restating of the choice.
```

C.2 Training Hyperparameters

Our SFT and DPO pipelines fine-tune `meta-llama/Llama-3.1-8B-Instruct` (8B parameters; Grattafiori et al., 2024) on ConFiQA. Unless otherwise specified, training runs use 3 epochs; the Llama SFT modified-only and multi-task runs use 1 epoch, since we observed overfitting within the first epoch. For DPO, we use the TRL `DPOTrainer` with a frozen reference model and the same chat-template setup mentioned above. DPO training uses per-device batch size 8, gradient accumulation 1, 2 GPUs per run (effective global batch size 16), warmup steps 100, cosine learning-rate schedule, `adamw_torch`, `bf16` training, gradient checkpointing enabled, maximum gradient norm 1.0, and seed 42. The DPO learning rate is `5e-7`, the maximum sequence length is 1024 tokens, and the maximum prompt length is 512 tokens. We use DeepSpeed ZeRO-3 for distributed training. For full SFT, we use an effective batch size of 64, warmup ratio 0.03, cosine learning-rate schedule, `adamw_torch`, `bf16` training, gradient checkpointing enabled, maximum gradient norm 1.0, seed 42, learning rate `2e-5`, and sequence length 1088 tokens. Hyperparameters were not formally swept. Our preference data and evaluation protocol are taken from Context-DPO (Bi et al., 2025), and our DPO maximum sequence and prompt lengths (1024 and 512 tokens) follow its released training code; the remaining DPO and SFT settings are standard choices that we did not tune.

Compute infrastructure. All training and evaluation was run on internal university clusters using NVIDIA L40S (48GB) and A100 (80GB) GPUs. SFT runs used 8 GPUs per run with FSDP `full_shard auto_wrap`, and DPO runs used 2 GPUs per run with DeepSpeed ZeRO-3. Aggregating across all training runs reported in this paper (modified-only, multi-task, and continual variants for SFT and DPO), the total training compute is approximately 300 GPU-hours. Inference for the cross-benchmark evaluation suite adds roughly an additional 100 GPU-hours, dominated by long-form generation on the held-out conflict benchmarks.

C.3 Reproducibility Note on Context-DPO Train-Test Splits

As the authors of Context-DPO (ConfiQA) did not release the exact training and evaluation splits used in their experiments, we reconstruct train-test splits by sampling the constituent splits as described in their previous publication. We use the public ConFiQA splits available through the HuggingFace release and fix the random seed to 42 for shuffling and sampling. ConFiQA consists of three sub-datasets (QA, MR, MC), each with 6000 instances. We sample 4500 instances per sub-dataset for training (13,500 total) and reserve the remaining 1500 per sub-dataset (4,500 total) for testing, maintaining a balanced distribution across hop levels (2-hop, 3-hop, 4-hop) within the MR and MC splits.

D Evaluation Protocol

Generation. All benchmarks use the prompts in Appendix C, greedy decoding, and at most 64 new tokens. Models answer directly, without chain-of-thought, including on GSM8K and BBH. Prompts longer than 2,048 tokens are skipped.

Context settings. For ConFiQA, ClashEval, HoH, and ConflictBank, each question is asked three ways: with the original context, scored against the original answer; with the modified context, scored against the modified answer; and without context, scored against the original answer. FaithEval has only the modified context. The closed-book benchmarks are asked without context.

Answer extraction and matching. For open-ended answers we take the first non-empty line after removing prefixes such as "Answer:". For multiple-choice questions we extract the option letter, and for GSM8K the final number. Predictions and gold answers are lowercased, punctuation and articles are removed, whitespace is collapsed, and number words and ordinals are mapped to digits. A prediction matches if any gold alias appears in it as a contiguous sequence of whole tokens.

Metrics. On ConFiQA we report the context-following rate Pc of Bi et al. (2025): a prediction counts only if it matches the target answer, does

not match the competing answer, and contains no negation word. Bi et al. (2025) define Pc for the modified context. We also apply it to the original-context and no-context settings, with the original answer as the target and the modified answer as the competitor, giving $Pc(m)$, $Pc(o)$, and $Pc(\text{no_ctx})$. All other benchmarks report accuracy: the percentage of examples where a gold answer alias appears in the normalized prediction as a contiguous sequence of whole tokens (for multiple choice, where the extracted option letter is correct).

Seeds. We evaluate every model with three seeds (42, 1, and 2 for Llama; 1, 2, and 3 for Qwen3). Each configuration is trained once, and decoding is greedy, so the seed only sets the order in which the evaluation data is shuffled. For benchmarks with more than 10,000 examples, this selects a different 10,000-example subset per seed. For smaller benchmarks it changes only the example order and batching, which causes small numerical differences. Fine-tuned models are compared with the base model on the same seed, and tables report the mean and sample standard deviation over seeds.

E Qwen3-8B Results

Table 4 reports the full Qwen3-8B results across all benchmarks, mirroring the layout of the Llama-3.1-8B-Instruct tables in the main text.

F Parametric Knowledge vs. Abstention

The no-context prompt permits `INSUFFICIENT_CONTEXT`, so a no-context drop can reflect abstention rather than forgetting. Table 5 reports no-context abstention rates. SFT variants rarely abstain (0.0% for every Llama SFT variant), so their no-context drops reflect forgetting. DPO variants abstain more than the Llama base model (e.g., PopQA 16.3% $\rightarrow$ 27.6–60.7%), so part of their no-context change reflects a more conservative abstention policy.

G Additional ConFiQA Results

We provide subset-level results for the constituent QA, MC, and MR splits in the ConFiQA dataset (See Table 6).

H Multi-task Ratio Definition

For SFT and DPO, multi-task ratios are reported as original-context:modified-context mixtures. A 1:1 ratio uses equal numbers of original-context and modified-context examples, while a 1:3 ratio uses all modified-context examples and samples one third as many original-context examples without replacement. Modified-only training corresponds to the no-original-context setting.

For DPO, the preferred and dispreferred responses swap across the two contexts: under the original context, the original answer is preferred and the counterfactual answer is rejected, while under the modified context, the counterfactual answer is preferred and the original answer is rejected. In both cases, the prompt text is formed from the corresponding context snippet and question, followed by the answer span.

In practice, we construct the final training set by including all modified-context examples and then adding a sampled subset of original-context examples according to the requested ratio, before shuffling the resulting mixture with a fixed random seed.

We denote continual-training variants with a `c` prefix. These are two-stage procedures rather than joint sampling: the model is first trained on original-context data and then continued on modified-context data. In our runs, `c1:1` means 3 epochs on original-context data followed by 3 epochs on modified-context data, while `c1:3` means 1 epoch on original-context data followed by 3 epochs on modified-context data.

Compute is not held fixed across settings. Because every augmented setting retains all modified-context examples and adds original-context examples on top, these runs are not compute-matched with modified-only training, and the multi-task and continual variants also differ from one another in total training budget. Table 7 lists, for each setting, the cumulative number of examples seen across all epochs and the resulting number of optimizer updates, for both SFT and DPO. Prompts longer than 2,048 tokens are skipped; for DPO both the chosen and rejected sequences must satisfy this limit, so the number of retained examples differs slightly between SFT and DPO. As discussed in the Limitations, we do not observe performance scaling with the training budget: position on the held-out Pareto frontier is governed by the mixture ratio rather than the number of examples seen or updates taken.

I AI Usage

During development we used Claude Code, Codex and GitHub Copilot for writing the training and

evaluation pipelines. While writing the paper, we used the same tools to debug LaTeX and to suggest rewrites of passages. The experimental design, results, and final text are our own.