COFFEQA: COUNTERFACTUAL-GUIDED FEW-SHOT DISTILLATION FOR LLM QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation aims to train a smaller student model to mimic a larger teacher model for resource-constrained environments. However, the process of distillation is typically data-intensive and computationally expensive, particularly for generative tasks. Few-shot distillation offers an alternative by transferring teacher knowledge using limited training samples. However, we observe that few-shot distillation performance is heavily dependent on the choice of these training samples. In fact, we demonstrate that using more uncertain samples near the teacher's decision boundary can significantly deteriorate distillation performance. Motivated by this observation, we propose ***Co**unter**f**actual-guided **Fe**w-shot Distillation in open-book **Q**uestion **A**nswering* (**COFFEQA**), a novel strategy that aptly uses counterfactuals to provide additional supervision around challenging factual samples. In the generative QA setting, we define a counterfactual as a minimally modified context that induces the teacher to change its response to an alternative answer. We provide geometric insights on how factual-counterfactual pairs can constrain the deviation between the teacher's and student's decision boundaries. We further introduce a new metric, ***C**ontext **F**ollowing **D**ifficulty* (**CFD**) that further validates counterfactuals prior to distillation based on agreement with the intended alternative answer and dependence on the modified context. COFFEQA uses CFD-validated factual-counterfactual pairs for few-shot distillation. Experiments across five in-domain and three out-of-domain QA benchmarks using T5 and Qwen2.5 families show that COFFEQA outperforms factual-only distillation baselines under the same few-shot budget. For instance, on TriviaQA, COFFEQA improves over the factual-only baseline by 8.77 average EM points and 9.98 average F1 points under the T5 setting with 32 few-shot pairs.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable success across a wide range of natural language processing tasks (Achiam et al., 2023; Chung et al., 2024; Gou et al., 2021; Raffel et al., 2020; Fang et al., 2026). However, their increasing size and computational requirements make deployment in resource-constrained environments particularly challenging (Kaplan et al., 2020; Brown et al., 2020; Strubell et al., 2019; Hoffmann et al., 2022). This motivates the need for compact models with low training and inference costs that preserve task-relevant capabilities of larger models (Liang et al., 2023; Saadi & Wang, 2025; Liu et al., 2024). Knowledge distillation (Hinton et al., 2015) seeks to address this challenge by transferring knowledge from a large teacher to a smaller student (Gou et al., 2021; Sanh et al., 2019; Xu et al., 2024b). However, most distillation approaches typically rely on large-scale datasets or extensive supervision from the teacher, thus incurring prohibitive training costs (Sanh et al., 2019; Sun et al., 2019; Jiao et al., 2020; Wang et al., 2020). This leads to our primary question: *How can we perform data-efficient distillation in LLMs?*

Towards addressing this question, our work studies few-shot distillation for generative tasks, with a focus on open-book question answering. We find that the performance of few-shot distillation is heavily dependent on the choice of the distillation dataset. For instance, samples near the teacher's decision boundary can significantly deteriorate distillation performance (see Fig. 6 in Section C), also consistent with prior findings on knowledge transfer around decision boundaries (Chen et al., 2025). This motivates us to provide additional supervision around these challenging regions while

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.

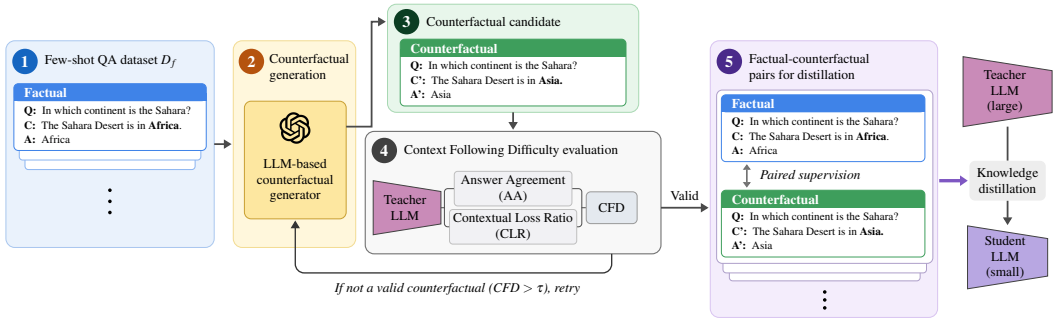


Figure 1: **Counterfactual-guided few-shot distillation framework.** We begin with (1) a few-shot factual QA dataset $\mathcal{D}_f$ and (2) generate a counterfactual candidate for a factual sample. (3) The generated counterfactual is evaluated using the teacher model with (4) our proposed Context Following Difficulty (CFD) metric and accepted if it satisfies the CFD threshold τ. (5) The CFD-validated counterfactuals are then paired with their corresponding factuals and used to distill the student model.

maintaining a limited few-shot budget. To achieve this, we leverage counterfactual examples. Counterfactual-based data augmentation has improved classification performance (Kaushik et al., 2019; Wu et al., 2021; Ross et al., 2021; Gardner et al., 2020; McAleese & Keane, 2024; Bhattacharjee et al., 2024) by encouraging models to capture causal dependencies and avoid shortcut learning (Guo et al., 2023; Chen et al., 2023b). However, its applicability to few-shot distillation for generative tasks remains largely unexplored. Unlike classification, generative tasks produce free-form outputs, making counterfactuals non-trivial to define. In generative tasks, we define a counterfactual to be a minimally modified input that changes the expected output while preserving overall semantic consistency, such as modifying supporting context in a question to induce a different answer (Guo et al., 2023; Longpre et al., 2021; Paranjape et al., 2022; Köksal et al., 2023; Yu et al., 2023).

In this work, we propose ***Co**unterf**a**ctual-guided **Fe**w-shot distillation in open-book **Q**uestion **A**nswering* (**CoFFeQA**), a novel strategy for few-shot distillation that aptly uses counterfactuals to provide additional supervision around challenging factual samples near the teacher's decision boundary. We demonstrate how selecting high-quality counterfactuals can guide knowledge transfer more effectively using extremely limited data. Our main contributions are as follows:

Counterfactual-guided few-shot distillation for generative tasks. We show that few-shot distillation becomes particularly challenging for factuals near the teacher's decision boundary, where the teacher has a smaller margin between its preferred answer and competing alternatives. Motivated by this observation, we propose **CoFFeQA**, a novel framework which leverages counterfactuals for few-shot distillation in open-book QA. Rather than relying solely on factual supervision, CoFFeQA pairs challenging factuals with their counterfactuals that expose changes in the teacher's response under minimal modifications to the input context, providing additional supervision around the teacher's decision boundary. This reduces dependence on large factual training corpora and improves training efficiency under limited supervision (see Figure 1 for an overview).

Geometric analysis of counterfactual-guided knowledge transfer. We first introduce a notion of decision boundaries for the teacher and student in the context-embedding space, motivated by the fact that variations in the input context can lead to different answers. Our formulation enables us to analyze distillation in open-book QA from a geometric perspective. Specifically, we show that factual-counterfactual pairs can act as probes for the student to approximate the local geometry of the teacher's decision boundary. To quantify the deviation between the teacher and student, we use the *Hausdorff distance* (Rockafellar & Wets, 1998), which measures the worst-case deviation between their decision boundaries. We show in Proposition 1 that this deviation is bounded by the magnitude of the counterfactual perturbations and their coverage of the teacher's decision boundary. Importantly, factual samples near the teacher's decision boundary can require smaller perturbations to induce a change in the teacher's response, which can lead to tighter teacher-student boundary alignment.

Context Following Difficulty (CFD) for counterfactual validation. Although we generate counterfactuals by perturbing the original input context to induce a targeted alternative answer,

not all generated examples are truly valid with respect to the teacher. The teacher may not produce same target answers using the modified context or may still be able to infer them without actually relying on the context. Such cases weaken the intended causal relationship that the student was intended to learn from the teacher and may encourage shortcut learning. To address this issue, we introduce a new metric *Context Following Difficulty (CFD)* to evaluate generated counterfactuals before distillation. CFD jointly considers whether the teacher produces the intended alternative answer and whether that answer becomes more predictable when conditioned on the modified context. COFFEQA uses CFD-validated counterfactuals, ensuring that the resulting supervision better captures the teacher's context-dependent behavior.

Empirical evaluation. We conduct comprehensive experiments across five in-domain and three out-of-domain benchmarks using the T5 (Raffel et al., 2020) and Qwen2.5 (Bai et al., 2023) model families. The in-domain benchmarks include NaturalQuestions, SQuAD, NewsQA, TriviaQA, and HotpotQA, while the out-of-domain benchmarks include DROP, DuoRC, and RelationExtraction (Fisch et al., 2019). We evaluate COFFEQA with standard knowledge distillation (KD), sequence-level knowledge distillation (SeqKD), and layer-wise distillation (LWD) baselines, under the same few-shot training budget. Across these settings, COFFEQA outperforms factual-only baselines. For instance, on TriviaQA, COFFEQA improves over the KD baseline by 8.77 average EM points and 9.98 average F1 points under the T5 setting with 32 few-shot pairs.

1.1 RELATED WORK

Counterfactual Data Augmentation in LLMs. Counterfactual data augmentation has been widely studied for improving model robustness and mitigating spurious correlations (Mishra et al., 2021; Hamman et al., 2024; Pawelczyk et al., 2020; Dissanayake & Dutta, 2024; Youssef et al., 2024; Kaushik et al., 2020; Egea et al., 2025; Sachdeva et al., 2024; Calderon et al., 2022). Early approaches relied on human-generated counterfactuals (Kaushik et al., 2019; Gardner et al., 2020), while later methods automated the generation (McAleese & Keane, 2024; Bhattacharjee et al., 2024), including LLM-based methods such as Polyjuice (Wu et al., 2021) and MiCE (Ross et al., 2021). Subsequent work has explored reinforcement learning (Chen et al., 2021), active learning (Balashankar et al., 2023), and word dictionary-based fairness-aware pipelines (Tokpo & Calders, 2024). Beyond classification, counterfactual methods have also been extended to generative tasks such as machine translation (Liu et al., 2021), dialogue (Ou et al., 2022), and summarization (Rajagopal et al., 2022). Unlike these works, which primarily use counterfactuals for robustness, reasoning, or data augmentation, we use factual-counterfactual pairs as structured supervision for knowledge distillation.

Counterfactual Question Answering. Counterfactual reasoning has recently gained traction in question answering (QA) (Ross et al., 2022; Khashabi et al., 2020; Huang et al., 2025). DisentQA (Neeman et al., 2023) uses counterfactual supervision to separate parametric and contextual knowledge (Xu et al., 2024a; Wang et al., 2023a), while IfQA (Yu et al., 2023) provides a benchmark for QA under hypothetical scenarios that require reasoning beyond factual retrieval. Counterfactual reasoning has also been studied in multi-hop QA (Guo et al., 2023) and CofCA (Wu et al., 2024), where counterfactual inference is used to mitigate shortcut reasoning. Retrieval-guided approaches further construct counterfactuals through context and answer modifications (Paranjape et al., 2022; Dixit et al., 2022). While these works demonstrate the value of counterfactuals for improving robustness and reasoning in QA, they do not study how they can facilitate knowledge distillation.

Distillation in LLMs. Knowledge Distillation (Hinton et al., 2015) aims to train smaller student models to mimic larger teacher models, with extensions to sequence-level distillation (Xu et al., 2024b; Kim & Rush, 2016; Gu et al., 2024; Yang et al., 2025), task-specific distillation (Sanh et al., 2019; Team et al., 2024; Wang et al., 2025), and layer-wise distillation (Jiao et al., 2020). While generative and instruction-level distillation often rely on direct supervision or synthetic data (Hsieh et al., 2023; Wang et al., 2023b; Gu et al., 2024), only a few recent works have incorporated counterfactuals into distillation (Hamman et al., 2025; Chen et al., 2023a). However, these approaches primarily focus on classification settings and do not address generative tasks, where outputs are open-ended and context-dependent. In contrast, we specifically focus on *few-shot* knowledge distillation for generative open-book QA. Our approach is motivated by the difficulty of distilling with factuals near the teacher's decision boundary and uses counterfactuals to provide additional supervision around these challenging regions. We also provide new geometric insights on counterfactual-guided knowledge transfer and a new metric CFD to further check counterfactual validity in open-book QA.

2 COFFEQA: COUNTERFACTUAL-GUIDED FEW-SHOT DISTILLATION

Our strategy is motivated by the observation that factual samples near the teacher's decision boundary are particularly challenging to distill under limited supervision. We use margin sampling (Scheffer et al., 2001) as an operational proxy for boundary proximity, where smaller teacher prediction margins indicate boundary near samples. Empirically, we observe that distillation performance degrades as this margin decreases; a detailed analysis is provided in Section C. Motivated by this observation, COFFEQA pairs boundary-near factuals with their counterfactuals that expose how the teacher's response changes under minimal variations of the input context. We first define and describe the counterfactual generation process, and then geometrically analyze how factual-counterfactual pairs facilitate teacher-student boundary alignment. Finally, we introduce Context Following Difficulty (CFD) to validate generated counterfactuals and present the overall COFFEQA distillation framework.

2.1 COUNTERFACTUAL DEFINITION AND GENERATION

We define counterfactual generation for open-book QA as constructing minimally modified training examples that preserve the original question while changing the evidence to support a different answer. Given a triplet (Q, C, A), where Q is the question, C is the context, and A is the answer, the counterfactual generation framework constructs a counterfactual by modifying the context using an alternative answer. The resulting counterfactual triplet is denoted as (Q, C', A'). We generate counterfactuals using GPT-4o-mini (OpenAI, 2024) in three stages. First, the answer A is classified into one of five entity types: *person, location, organization, date*, or *numeric*; samples outside these categories are excluded. Second, GPT-4o-mini generates an alternative answer A' of the same entity type that is similar in category but distinct from A. Previously generated alternatives are excluded to encourage diversity across generation attempts. Finally, GPT-4o-mini is provided with the original triplet (Q, C, A) and A' to generate the counterfactual context C' by minimally editing C. Specifically, occurrences of A are replaced with A', together with any related references that must be changed to keep the passage grammatically consistent. As a result, the modified context C' supports A' as the answer to the same question Q, yielding the counterfactual triplet (Q, C', A'). The system and user prompts used are shown in Figure 2. Additional prompt details are provided in Appendix D.

Counterfactual Context Generation: System Prompt

You are an expert at creating counterfactual reading-comprehension passages for QA training data. Given an original context that supports a specific answer, you rewrite it MINIMALLY so that it instead unambiguously supports a different target answer, by replacing occurrences of the original answer (and grammatically-dependent references to it) with the target answer, while leaving everything else in the passage unchanged. Do not add explanations or extra content.

Counterfactual Context Generation: User Prompt

Question: `{question}`, Original context (supports answer "`{original_answer}`"): `{original_context}`

Target answer (the new counterfactual answer): "`{target_answer}`"

Rewrite the context above so that it minimally changes to unambiguously support "`{target_answer}`" as the answer to the question, instead of "`{original_answer}`". Replace occurrences of "`{original_answer}`" (and any grammatically-dependent references to it) with "`{target_answer}`", keeping everything else in the passage exactly the same. Do not add commentary. Respond with ONLY: `{"reasoning": "<1-2 sentences on what you changed>", "counterfactual_context": "<the full rewritten context>"}`

Figure 2: **Counterfactual context generation prompts.** System and user prompts used to minimally modify the factual context C into a counterfactual context C' that supports the target answer A'.

2.2 GEOMETRIC ANALYSIS OF COUNTERFACTUAL-GUIDED KNOWLEDGE TRANSFER

We provide a geometric perspective on how factual-counterfactual pairs can facilitate knowledge transfer. We first introduce a notion of a decision boundary for open-book QA. While classification tasks typically consider a *label-flipping boundary*, we define a *context-response boundary* in the context-embedding space. In classification, the boundary separates regions where two classes have equal probabilities. In contrast, in open-book QA, it represents the transition region in context embedding space where the model shifts from generating one answer to another answer. In open-book QA, the context space may contain regions corresponding to the factual answer A, the intended counterfactual answer A', and other

answer (A'') regions, different from both A and A' corresponding to out-of-context or unrelated outputs. Our analysis focuses specifically on the boundary between the A and A' response regions, as this is the boundary relevant to factual-counterfactual supervision. Our CFD-based validation helps identify and reject counterfactual candidates that induce responses corresponding to the other answer (A'') regions before distillation. We therefore restrict the geometric analysis to the boundary between factual answer space (A) and intended counterfactual answer (A') space. We use the Hausdorff distance (Rockafellar & Wets, 1998) to quantify the worst-case deviation between the teacher context-response boundary and the student context-response boundary and derive an upper bound for this deviation in Proposition 1 (The proof can be found in Appendix E).

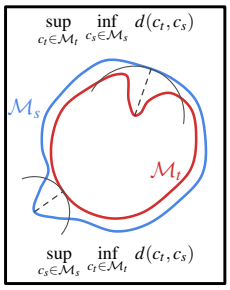


Figure 3: Visualizing the Hausdorff distance between $\mathcal{M}_t$ and $\mathcal{M}_s$.

Definition of Hausdorff Distance (Rockafellar & Wets, 1998). Let $(\mathcal{M}, d)$ be a metric space. For each pair of non-empty subsets $\mathcal{M}_t \subset \mathcal{M}$ and $\mathcal{M}_s \subset \mathcal{M}$, the Hausdorff distance between $\mathcal{M}_t$ and $\mathcal{M}_s$ is defined as:

$$d_H(\mathcal{M}_t, \mathcal{M}_s) := \max\left\{ \sup_{c_t \in \mathcal{M}_t} \inf_{c_s \in \mathcal{M}_s} d(c_t, c_s), \sup_{c_s \in \mathcal{M}_s} \inf_{c_t \in \mathcal{M}_t} d(c_t, c_s) \right\}.$$

This measures the maximum of the minimum distances between two sets, capturing the worst-case geometric deviation. (see Figure 3).

Proposition 1 *Let $\mathcal{C}$ be the space of context embeddings, $c \in \mathcal{C}$ arbitrary input context, $Q \in \mathcal{Q}$ input question, $A \in \mathcal{A}$ factual answer, and $A' \in \mathcal{A}'$ counterfactual answer. Consider teacher and student models with response functions $G_t(Q, c) = \arg\max_A P_t(A|Q, c)$ and $G_s(Q, c) = \arg\max_A P_s(A|Q, c)$. The respective context-flipping boundaries can be defined as the sets:*

$$\mathcal{M}_t = \{c \in \mathcal{C} \mid P_t(A|Q, c) = P_t(A'|Q, c)\}, \quad \mathcal{M}_s = \{c \in \mathcal{C} \mid P_s(A|Q, c) = P_s(A'|Q, c)\}.$$

Suppose there exists a factual-counterfactual paired dataset $\mathcal{D}_{f+cf} = \{(c_i, c_i')\}_{i=1}^k$ for the above $\mathcal{M}_t$ that satisfies three conditions: (1) Minimal context perturbation in context embedding space is bounded by α (i.e., $||c_i - c_i'|| \leq \alpha$), (2) Exact answer generation: $G_t(Q, c) = G_s(Q, c)$ where $c = \{c_i, c_i'\}$, (3) ϵ-spread boundary coverage: Counterfactuals are well-distributed across the context-response boundaries with spread ϵ, i.e., if $t \in \mathcal{M}_t$, then $\exists t_i^*$ such that $\|t - t_i^*\| \leq \epsilon$. This applies for $\mathcal{M}_s$ as well. i.e., if $s \in \mathcal{M}_s$, then $\exists s_i^*$ such that $\|s - s_i^*\| \leq \epsilon$. Then, the Hausdorff distance between the teacher and student boundaries satisfies: $d_H(\mathcal{M}_t, \mathcal{M}_s) \leq \alpha + \epsilon$.

Intuition behind proposition and assumptions. Proposition 1 highlights two properties of factual-counterfactual supervision that govern teacher-student boundary alignment: the magnitude of the counterfactual perturbations (α) and the coverage of the context-response boundary (ϵ). First, each factual-counterfactual pair provides supervision across a local change in the teacher's response. Importantly, when the factual sample is already close to the teacher's decision boundary, a smaller contextual perturbation is sufficient to cross the boundary and induce the alternative response A'. Therefore, starting from boundary-near factual samples leads to a smaller upper bound on the perturbation α, tightening the bound. Although those samples are difficult under factual-only supervision, their proximity to the boundary makes them useful starting points for constructing local counterfactual supervision. Second, the boundary coverage condition captures how well the induced crossing points cover both the context-response boundaries. Better coverage corresponds to a smaller ϵ, tightening the bound. Thus, we provide a geometric characterization of how local counterfactual perturbations and boundary coverage relate to teacher-student boundary deviation.

The analysis above also assumes that the student agrees with the teacher on the paired factual and counterfactual samples (Assumption (2) Exact answer generation). This idealized condition captures the objective of distillation, where the student is trained to reproduce the teacher's behavior on the available supervision. More importantly, the analysis also relies on the generated counterfactual context C' to induce the intended alternative answer A'. In practice, this may not be guaranteed due to inherently stochastic and unpredictable generative behavior of LLMs Holtzman et al. (2019). The teacher may fail to produce A', or conditioning on C' may provide only limited additional predictive benefit for A'. For realization of this objective in our implementation, we therefore introduce a new metric *Context Following Difficulty (CFD)* in the following section to validate generated counterfactuals before they are used for distillation.

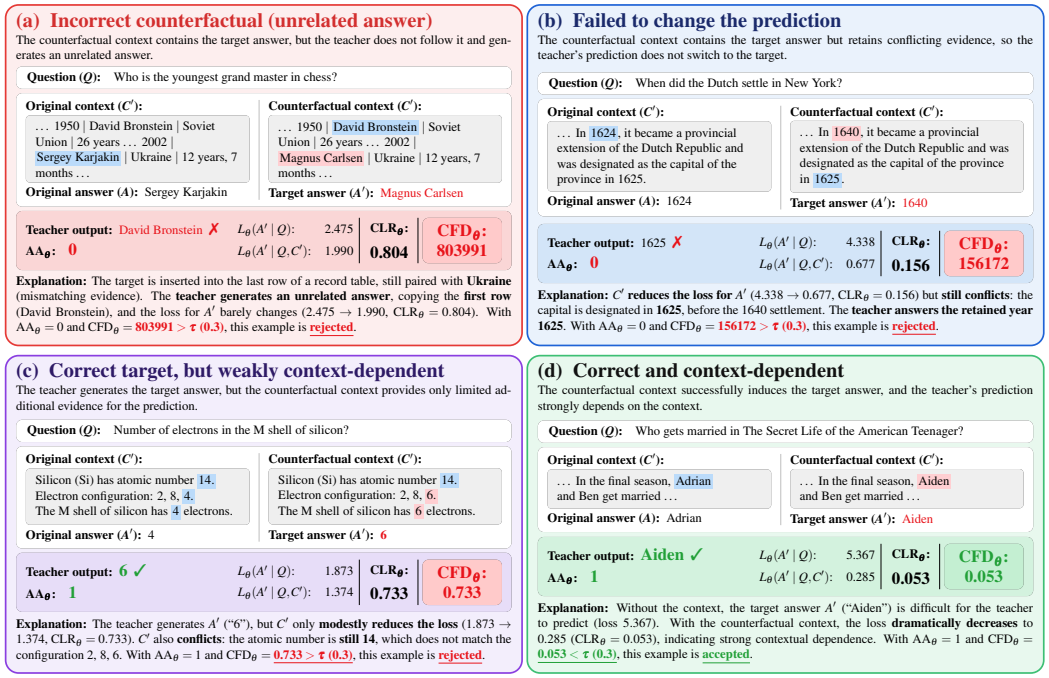


Figure 4: **Illustrative examples of CFD-based counterfactual evaluation from the NQ dataset.** The four cases demonstrate how Answer Agreement (AA) and Contextual Loss Ratio (CLR) complement each other in evaluating generated counterfactual candidates.

2.3 Context Following Difficulty (CFD) for Counterfactual Validation

Generated counterfactual samples should be evaluated using the teacher to determine whether they induce the intended alternative answer A' rather than another response. Since exact matching can be unreliable for generative tasks (Holtzman et al., 2019), we use token-level F1 (Rajpurkar et al., 2016) to measure agreement between the teacher output and A'. However, answer agreement alone is insufficient, as generating the intended answer does not necessarily indicate that the C' provides strong evidence for that prediction. A high-quality counterfactual should therefore not only induce the intended answer A', but also make A' substantially more predictable when conditioned on C' than from the Q alone. To capture these two properties, we introduce a new metric *Context Following Difficulty (CFD)*, which combines *Answer Agreement (AA)* with the *Contextual Loss Ratio (CLR)*. First, AA_θ measures whether the teacher generates the intended alternative answer A' under the counterfactual context C':

$$\mathrm{AA}_\theta(Q, C', A') = \mathrm{F1}\big(\hat{A}_\theta(Q, C'), A'\big), \qquad (1)$$

where $\hat{A}_\theta(Q, C')$ is the answer generated by the teacher given Q and C'. A higher AA_θ indicates stronger agreement with the intended answer A'. However, high answer agreement alone does not establish that the counterfactual context plays a substantial role in predicting A'. Inspired by the Instruction Following Difficulty (IFD) metric (Li et al., 2024), we define the *Contextual Loss Ratio (CLR)* as:

$$\mathrm{CLR}_\theta(Q, C', A') = \frac{L_\theta(A' \mid Q, C')}{L_\theta(A' \mid Q)}. \qquad (2)$$

Algorithm 1 CoFFeQA

Input: $\mathcal{D}$, teacher θ_t, student θ_s, epochs E, learning rate η, weight α, CFD threshold τ, few-shot budget k, $\mathcal{D}_{cf}=\phi, \mathcal{D}_f=\phi$

while $|\mathcal{D}_{cf}| < k/2$ **do**
  Pick $(Q, C, A) \in \mathcal{D}$
  $(Q, C', A') \leftarrow \mathrm{GenCF}(Q, C, A)$
  if $\mathrm{CFD}_{\theta_t}(Q, C', A') \le \tau$ **then**
    $\mathcal{D}_{cf} \leftarrow \mathcal{D}_{cf} \cup \{(Q, C', A')\}$
    $\mathcal{D}_f \leftarrow \mathcal{D}_f \cup \{(Q, C, A)\}$
  end if
end while
for $e = 1$ **to** E **do**
  for all $(Q, C, A) \in \mathcal{D}_f \cup \mathcal{D}_{cf}$ **do**
    $\mathcal{L}_{\mathrm{CE}} \leftarrow \mathrm{CE}(\theta_s(Q, C), A)$
    $\mathcal{L}_{\mathrm{KD}} \leftarrow \mathrm{KL}(\theta_t(Q, C) \| \theta_s(Q, C))$
    $\mathcal{L} \leftarrow \alpha\mathcal{L}_{\mathrm{CE}} + (1-\alpha)\mathcal{L}_{\mathrm{KD}}$
    Update $\theta_s \leftarrow \theta_s - \eta\nabla_{\theta_s}\mathcal{L}$
  end for
end for
return distilled student θ_s

A low CLR_θ indicates that conditioning on C' substantially reduces the teacher's loss for A' relative to conditioning on Q alone, i.e., C' plays a strong role in predicting the intended alternative answer. Conversely, a higher CLR_θ indicates that providing C' yields only a limited reduction in the loss of A', suggesting weaker dependence of the target prediction on C'. Thus, CLR_θ measures how strongly C' contributes to the teacher's prediction of A'.

Combining these two measures, we finally define our *Context Following Difficulty (CFD)* score as:

$$\mathrm{CFD}_\theta(Q, C', A') = \frac{\mathrm{CLR}_\theta(Q, C', A')}{\mathrm{AA}_\theta(Q, C', A') + \delta}, \tag{3}$$

where $\delta > 0$ is a small constant for numerical stability. We prefer low CLR_θ and high AA_θ. A lower CFD_θ score favors counterfactuals for which the teacher both generates an answer close to the intended target A' and relies more strongly on the counterfactual context when predicting A'. Figure 4 illustrates the complementary roles of AA_θ and CLR_θ in four representative cases. Cases (a) and (b) show that a counterfactual may fail to induce the intended answer even when the modified context contains or supports it. Case (c) has high AA_θ but high CLR_θ, indicating limited additional predictive benefit from C'. In contrast, case (c) represents a valid counterfactual with a low CFD_θ. Together, these examples show why both AA_θ and CLR_θ are necessary for validation. The two components of CFD_θ therefore capture complementary properties of a valid and informative counterfactual. AA_θ verifies the intended answer A', while CLR_θ measures how strongly the modified context contributes to predicting A'. Our complete algorithm is provided in Algorithm 1 (also in Figure 1).

3 EXPERIMENTS

Datasets and Models. We evaluate COFFEQA on eight open-domain question answering benchmarks. The five in-domain benchmarks are Natural Questions (NQ) (Kwiatkowski et al., 2019), SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), and HotpotQA (Yang et al., 2018). The three out-of-domain benchmarks are DROP (Dua et al., 2019), DuoRC (Saha et al., 2018), and RelationExtraction (Levy et al., 2017). We evaluate our method using two model families with different architectures: the encoder-decoder model T5 (Raffel et al., 2020) and the decoder-only model Qwen2.5 (Bai et al., 2023). In the T5 setting, T5-Large (770M parameters) serves as the teacher, while T5-Base (220M parameters) serves as the student. In the Qwen2.5 setting, Qwen2.5-7B-IT serves as the teacher, while Qwen2.5-1.5B-IT serves as the student. Additional details are provided in Appendix F.1 (datasets) and Appendix F.2 (full-training).

Baselines. We compare COFFEQA against three distillation baselines: (i) Standard Knowledge Distillation (KD) (Hinton et al., 2015), (ii) Sequence-Level Knowledge Distillation (seqKD) (Kim & Rush, 2016), and (iii) Layer-Wise Distillation (LWD) (Jiao et al., 2020). All methods use the same few-shot splits for fair comparison. Additional details are provided in Appendix F.3.

Setup. Following prior work (Hamman et al., 2025), we first fine-tune the teacher model on the full training dataset. For each few-shot budget $k \in \{16, 32, 64, 128, 256\}$, we rank factual samples by their teacher prediction margins, using margin sampling (Scheffer et al., 2001). The factual-only distillation baseline uses the k samples with the smallest margins. COFFEQA instead uses the $k/2$ factual samples with the smallest margins together with their corresponding $k/2$ counterfactuals, maintaining the same total few-shot budget k. We evaluate accuracy using EM and F1 against the gold answers and teacher-student agreement using EM and F1 between their generated answers. All results are reported as the mean and standard deviation over five random seeds.

3.1 MAIN EXPERIMENTAL RESULTS

COFFEQA outperforms the baselines. As shown in Table 1, COFFEQA improves performance across all three distillation baselines under the T5 setting. When averaged across all datasets and k-shot settings, COFFEQA improves the F1 score from 66.42 to 69.27 with standard KD, from 66.37 to 69.28 with LWD, and from 68.63 to 70.52 with SeqKD. Moreover, COFFEQA exhibits lower standard deviations than the corresponding baselines in many settings, indicating more stable performance across seeds. These results demonstrate that counterfactual supervision provides consistent benefits over the baselines. Detailed EM scores are provided in Table 8 in Appendix H.1.

Table 1: In-domain F1 score comparison across k-shots for T5 and Qwen2.5 models using baselines. Mean F1 scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. baseline). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$61.83_{0.49}$	$61.44_{2.08}$	$60.34_{1.91}$	$56.91_{3.20}$	$58.07_{1.33}$	$59.72_{2.15}$	$\mathbf{66.91}_{0.44}$	$66.01_{0.59}$	$65.71_{0.39}$	$64.68_{0.61}$	$65.66_{0.69}$	$65.79_{0.80}$
	+ CoFFeQA	$\mathbf{62.16}_{0.27}$	$\mathbf{62.09}_{1.30}$	$\mathbf{60.85}_{1.37}$	$\mathbf{59.08}_{2.01}$	$\mathbf{58.63}_{2.30}$	$\mathbf{60.56}_{1.65}$	$66.27_{0.27}$	$\mathbf{68.10}_{0.20}$	$\mathbf{67.23}_{0.32}$	$\mathbf{67.25}_{0.73}$	$\mathbf{67.19}_{0.99}$	$\mathbf{67.21}_{0.65}$
	LWD	$61.81_{0.56}$	$61.23_{2.49}$	$\mathbf{60.37}_{1.56}$	$58.16_{1.81}$	$58.72_{2.29}$	$60.06_{1.57}$	$\mathbf{65.62}_{5.03}$	$66.32_{0.94}$	$65.90_{1.18}$	$65.61_{0.85}$	$66.62_{0.86}$	$66.01_{0.45}$
	+ CoFFeQA	$\mathbf{62.12}_{0.22}$	$\mathbf{61.46}_{2.57}$	$60.04_{2.09}$	$\mathbf{60.03}_{1.54}$	$\mathbf{58.93}_{2.71}$	$\mathbf{60.52}_{1.27}$	$63.98_{4.96}$	$\mathbf{68.16}_{0.35}$	$\mathbf{67.78}_{0.64}$	$\mathbf{68.07}_{0.56}$	$\mathbf{67.96}_{0.77}$	$\mathbf{67.19}_{1.80}$
	SeqKD	$60.68_{1.24}$	$59.19_{1.07}$	$60.80_{1.12}$	$59.78_{1.57}$	$\mathbf{62.09}_{1.54}$	$60.51_{1.11}$	$65.89_{0.34}$	$\mathbf{67.55}_{0.55}$	$66.84_{0.31}$	$67.33_{0.35}$	$67.73_{0.76}$	$67.07_{0.74}$
	+ CoFFeQA	$\mathbf{61.51}_{0.69}$	$\mathbf{61.88}_{0.64}$	$\mathbf{61.52}_{0.92}$	$\mathbf{61.87}_{0.81}$	$61.36_{1.41}$	$\mathbf{61.62}_{0.23}$	$\mathbf{66.56}_{0.42}$	$66.92_{0.74}$	$\mathbf{68.48}_{0.89}$	$\mathbf{68.04}_{0.58}$	$\mathbf{68.22}_{0.39}$	$\mathbf{67.64}_{0.85}$
SQuAD	KD	$86.45_{0.54}$	$84.18_{1.39}$	$83.63_{1.71}$	$84.05_{1.14}$	$83.10_{1.19}$	$84.28_{1.28}$	$\mathbf{77.75}_{0.13}$	$76.78_{0.24}$	$76.40_{0.47}$	$75.87_{0.41}$	$\mathbf{76.45}_{0.39}$	$76.65_{0.70}$
	+ CoFFeQA	$\mathbf{86.56}_{0.16}$	$\mathbf{86.13}_{0.72}$	$\mathbf{85.36}_{1.03}$	$\mathbf{85.46}_{0.48}$	$\mathbf{84.23}_{1.58}$	$\mathbf{85.55}_{0.89}$	$76.95_{0.46}$	$\mathbf{77.28}_{0.59}$	$\mathbf{77.42}_{0.42}$	$\mathbf{76.64}_{0.85}$	$75.95_{0.34}$	$\mathbf{76.85}_{0.59}$
	LWD	$86.47_{0.50}$	$84.13_{1.13}$	$83.86_{1.78}$	$83.45_{0.83}$	$82.92_{1.18}$	$84.17_{1.37}$	$\mathbf{76.70}_{0.65}$	$76.96_{0.30}$	$76.71_{0.41}$	$76.15_{1.00}$	$77.20_{1.07}$	$76.74_{0.39}$
	+ CoFFeQA	$\mathbf{86.56}_{0.23}$	$\mathbf{86.09}_{0.78}$	$\mathbf{84.85}_{0.81}$	$\mathbf{85.57}_{0.49}$	$\mathbf{84.53}_{1.65}$	$\mathbf{85.52}_{0.84}$	$76.36_{0.06}$	$\mathbf{77.19}_{0.24}$	$\mathbf{77.75}_{0.41}$	$\mathbf{77.22}_{0.33}$	$\mathbf{77.82}_{1.50}$	$\mathbf{77.27}_{0.58}$
	SeqKD	$85.43_{0.69}$	$\mathbf{86.25}_{0.67}$	$85.57_{0.28}$	$85.45_{0.32}$	$84.96_{0.52}$	$85.53_{0.46}$	$\mathbf{78.22}_{0.46}$	$77.79_{0.17}$	$\mathbf{77.90}_{0.51}$	$77.35_{0.49}$	$78.26_{0.96}$	$77.90_{0.37}$
	+ CoFFeQA	$\mathbf{86.41}_{0.45}$	$85.47_{0.51}$	$\mathbf{85.73}_{0.73}$	$\mathbf{85.65}_{0.36}$	$\mathbf{85.02}_{0.37}$	$\mathbf{85.66}_{0.50}$	$77.25_{0.37}$	$\mathbf{77.89}_{0.33}$	$77.80_{0.65}$	$\mathbf{77.85}_{0.53}$	$\mathbf{78.63}_{0.67}$	$\mathbf{77.92}_{0.49}$
NewsQA	KD	$72.72_{2.33}$	$73.43_{1.92}$	$70.56_{2.16}$	$69.95_{2.11}$	$65.35_{2.07}$	$70.40_{3.17}$	$57.32_{0.77}$	$54.14_{1.48}$	$54.32_{1.11}$	$55.10_{1.69}$	$53.64_{1.71}$	$54.90_{1.45}$
	+ CoFFeQA	$\mathbf{73.66}_{3.02}$	$\mathbf{74.26}_{1.18}$	$\mathbf{75.43}_{1.71}$	$\mathbf{74.19}_{1.38}$	$\mathbf{72.85}_{2.24}$	$\mathbf{74.08}_{0.94}$	$\mathbf{60.77}_{0.73}$	$\mathbf{57.95}_{0.82}$	$\mathbf{56.88}_{0.60}$	$\mathbf{57.48}_{1.17}$	$\mathbf{59.41}_{2.54}$	$\mathbf{58.50}_{1.58}$
	LWD	$72.67_{2.30}$	$73.00_{2.05}$	$70.42_{2.59}$	$69.29_{2.49}$	$65.25_{0.79}$	$70.12_{3.14}$	$58.76_{1.01}$	$\mathbf{58.62}_{0.68}$	$56.90_{0.93}$	$57.51_{1.55}$	$55.72_{2.27}$	$57.50_{1.26}$
	+ CoFFeQA	$\mathbf{73.43}_{3.03}$	$\mathbf{74.98}_{0.66}$	$\mathbf{74.85}_{1.56}$	$\mathbf{73.77}_{1.18}$	$\mathbf{73.06}_{2.36}$	$\mathbf{74.02}_{0.86}$	$\mathbf{60.49}_{0.88}$	$58.18_{1.05}$	$\mathbf{58.05}_{1.13}$	$\mathbf{59.89}_{0.48}$	$\mathbf{61.23}_{4.39}$	$\mathbf{59.57}_{1.41}$
	SeqKD	$66.77_{1.63}$	$70.09_{1.72}$	$70.21_{3.11}$	$69.45_{1.83}$	$72.16_{0.81}$	$69.74_{1.94}$	$49.10_{1.20}$	$51.49_{1.04}$	$57.60_{0.49}$	$58.97_{1.25}$	$61.61_{1.31}$	$55.75_{5.25}$
	+ CoFFeQA	$\mathbf{72.72}_{1.51}$	$\mathbf{74.65}_{1.26}$	$\mathbf{75.99}_{1.48}$	$\mathbf{76.54}_{1.88}$	$\mathbf{75.77}_{1.46}$	$\mathbf{75.13}_{1.51}$	$\mathbf{51.22}_{1.13}$	$\mathbf{55.65}_{0.88}$	$\mathbf{56.90}_{0.95}$	$\mathbf{60.14}_{1.62}$	$\mathbf{63.38}_{1.95}$	$\mathbf{57.46}_{4.60}$
TriviaQA	KD	$53.70_{4.24}$	$50.40_{3.27}$	$46.49_{2.23}$	$45.80_{2.86}$	$46.93_{1.44}$	$48.66_{3.33}$	$61.57_{1.09}$	$\mathbf{61.98}_{0.79}$	$59.92_{1.06}$	$57.50_{1.14}$	$\mathbf{59.26}_{1.01}$	$60.05_{1.82}$
	+ CoFFeQA	$\mathbf{60.06}_{1.45}$	$\mathbf{56.32}_{2.19}$	$\mathbf{56.47}_{7.04}$	$\mathbf{50.54}_{1.92}$	$\mathbf{49.52}_{1.77}$	$\mathbf{54.58}_{4.43}$	$\mathbf{62.49}_{0.42}$	$60.58_{0.69}$	$\mathbf{62.32}_{0.67}$	$\mathbf{60.53}_{0.74}$	$57.96_{1.49}$	$\mathbf{60.77}_{1.83}$
	LWD	$53.73_{3.89}$	$51.86_{2.61}$	$44.61_{2.48}$	$45.08_{1.90}$	$47.19_{1.42}$	$48.50_{4.10}$	$61.10_{0.41}$	$61.20_{1.11}$	$59.26_{0.40}$	$58.07_{0.42}$	$\mathbf{58.86}_{1.29}$	$59.70_{1.40}$
	+ CoFFeQA	$\mathbf{59.70}_{1.51}$	$\mathbf{56.49}_{2.51}$	$\mathbf{56.35}_{7.32}$	$\mathbf{50.46}_{1.88}$	$\mathbf{49.67}_{1.09}$	$\mathbf{54.53}_{4.30}$	$\mathbf{62.25}_{0.68}$	$\mathbf{61.49}_{0.79}$	$\mathbf{61.76}_{0.50}$	$\mathbf{60.21}_{1.16}$	$57.87_{1.39}$	$\mathbf{60.72}_{1.76}$
	SeqKD	$59.01_{3.46}$	$52.99_{1.83}$	$50.19_{2.62}$	$50.53_{2.14}$	$48.61_{1.73}$	$52.27_{4.08}$	$61.36_{1.02}$	$59.99_{0.54}$	$60.52_{0.81}$	$59.85_{1.07}$	$\mathbf{60.58}_{1.30}$	$60.46_{0.60}$
	+ CoFFeQA	$\mathbf{60.70}_{2.14}$	$\mathbf{55.11}_{1.75}$	$\mathbf{54.79}_{1.59}$	$\mathbf{51.05}_{3.29}$	$\mathbf{53.40}_{1.24}$	$\mathbf{55.01}_{3.56}$	$\mathbf{63.13}_{0.84}$	$\mathbf{62.05}_{1.05}$	$\mathbf{61.76}_{0.76}$	$\mathbf{60.17}_{0.64}$	$59.86_{0.66}$	$\mathbf{61.39}_{1.36}$
HotpotQA	KD	$\mathbf{70.37}_{2.43}$	$70.96_{3.32}$	$64.27_{3.26}$	$70.23_{1.10}$	$69.42_{3.49}$	$69.05_{2.73}$	$\mathbf{74.09}_{0.48}$	$\mathbf{75.35}_{0.53}$	$72.73_{1.19}$	$73.15_{0.57}$	$73.10_{0.97}$	$73.68_{1.06}$
	+ CoFFeQA	$69.87_{3.08}$	$\mathbf{73.39}_{2.01}$	$\mathbf{71.31}_{1.26}$	$\mathbf{70.65}_{1.90}$	$\mathbf{72.61}_{2.43}$	$\mathbf{71.56}_{1.43}$	$73.38_{0.92}$	$74.79_{0.50}$	$\mathbf{76.44}_{0.65}$	$\mathbf{75.69}_{0.61}$	$\mathbf{76.17}_{1.07}$	$\mathbf{75.29}_{1.24}$
	LWD	$\mathbf{69.89}_{2.41}$	$70.77_{3.76}$	$64.11_{3.08}$	$69.88_{1.68}$	$70.44_{2.77}$	$69.02_{2.77}$	$\mathbf{74.20}_{0.49}$	$74.45_{0.78}$	$72.40_{0.86}$	$74.36_{0.54}$	$74.07_{1.84}$	$73.90_{0.85}$
	+ CoFFeQA	$69.21_{3.32}$	$\mathbf{74.09}_{1.82}$	$\mathbf{72.04}_{1.56}$	$\mathbf{70.50}_{2.17}$	$\mathbf{73.28}_{2.23}$	$\mathbf{71.82}_{1.99}$	$73.87_{0.91}$	$\mathbf{74.77}_{0.41}$	$\mathbf{74.70}_{0.53}$	$\mathbf{75.75}_{0.89}$	$\mathbf{77.32}_{0.82}$	$\mathbf{75.28}_{1.32}$
	SeqKD	$\mathbf{75.81}_{0.84}$	$75.65_{0.88}$	$75.79_{1.16}$	$74.16_{2.96}$	$\mathbf{74.22}_{3.15}$	$75.13_{0.86}$	$\mathbf{75.37}_{0.66}$	$\mathbf{76.49}_{0.78}$	$75.26_{0.50}$	$72.29_{0.71}$	$73.59_{1.65}$	$74.60_{1.66}$
	+ CoFFeQA	$74.84_{0.98}$	$\mathbf{76.99}_{0.54}$	$\mathbf{75.84}_{1.63}$	$\mathbf{75.03}_{2.36}$	$73.14_{3.32}$	$\mathbf{75.17}_{1.41}$	$73.67_{0.66}$	$74.07_{1.30}$	$\mathbf{76.23}_{0.70}$	$\mathbf{76.04}_{0.56}$	$\mathbf{74.90}_{1.46}$	$\mathbf{74.98}_{1.14}$
AVG	KD	$69.01_{12.31}$	$68.08_{12.77}$	$65.06_{13.63}$	$65.39_{14.56}$	$64.58_{13.42}$	$66.42_{13.25}$	$67.53_{8.48}$	$66.85_{9.44}$	$65.81_{9.04}$	$65.26_{9.20}$	$65.62_{9.45}$	$66.21_{9.09}$
	+ CoFFeQA	$\mathbf{70.46}_{10.57}$	$\mathbf{70.44}_{11.60}$	$\mathbf{69.88}_{11.55}$	$\mathbf{67.98}_{13.55}$	$\mathbf{67.57}_{13.57}$	$\mathbf{69.27}_{12.09}$	$\mathbf{67.97}_{6.97}$	$\mathbf{67.74}_{8.48}$	$\mathbf{68.06}_{8.89}$	$\mathbf{67.52}_{8.66}$	$\mathbf{67.34}_{8.70}$	$\mathbf{67.72}_{8.28}$
	LWD	$68.91_{12.29}$	$68.20_{12.24}$	$64.67_{14.34}$	$65.17_{14.37}$	$64.90_{13.30}$	$66.37_{13.20}$	$67.28_{7.91}$	$67.51_{8.03}$	$66.24_{8.42}$	$66.34_{8.77}$	$66.49_{9.31}$	$66.77_{8.47}$
	+ CoFFeQA	$\mathbf{70.21}_{10.66}$	$\mathbf{70.62}_{11.77}$	$\mathbf{69.63}_{11.55}$	$\mathbf{68.07}_{13.41}$	$\mathbf{67.89}_{13.65}$	$\mathbf{69.28}_{12.11}$	$\mathbf{67.39}_{7.21}$	$\mathbf{67.96}_{8.20}$	$\mathbf{68.01}_{8.34}$	$\mathbf{68.23}_{8.24}$	$\mathbf{68.44}_{9.09}$	$\mathbf{68.00}_{8.12}$
	SeqKD	$69.54_{11.05}$	$68.84_{13.18}$	$68.51_{13.62}$	$67.87_{13.39}$	$68.41_{13.73}$	$68.63_{12.87}$	$65.99_{11.67}$	$66.66_{11.13}$	$67.62_{8.88}$	$67.16_{7.92}$	$68.35_{7.61}$	$67.16_{9.30}$
	+ CoFFeQA	$\mathbf{71.24}_{10.62}$	$\mathbf{70.82}_{12.19}$	$\mathbf{70.77}_{12.42}$	$\mathbf{70.03}_{13.59}$	$\mathbf{69.74}_{12.44}$	$\mathbf{70.52}_{12.16}$	$\mathbf{66.37}_{10.14}$	$\mathbf{67.32}_{8.96}$	$\mathbf{68.27}_{9.08}$	$\mathbf{68.45}_{8.42}$	$\mathbf{69.00}_{7.80}$	$\mathbf{67.88}_{8.69}$

CoFFeQA generalizes across different model families. CoFFeQA improves standard KD across both the encoder-decoder T5 and decoder-only Qwen2.5 model families, where the teachers are approximately $3.5\times$ and $4.7\times$ larger than their respective students. Under Qwen2.5, CoFFeQA improves the average F1 score from 66.21 to 67.72 with standard KD, from 66.77 to 68.00 with LWD, and from 67.16 to 67.88 with SeqKD. These results indicate that the benefits of CoFFeQA extend across different model architectures and scales.

Students mimic the teacher more closely under CoFFeQA. To examine how closely the student follows the teacher, we measure teacher-student agreement using EM and F1. As shown in Figure 5, CoFFeQA generally achieves higher teacher-student agreement than standard KD across datasets and k-shot settings. Averaged across all benchmarks and settings, CoFFeQA improves F1 agreement under standard KD from 68.51 to 71.50 under T5, with a similar trend for Qwen2.5. These results suggest that counterfactual supervision helps the student more closely reproduce the teacher's output behavior. Detailed results and additional radar plots are provided in Appendix H.2.

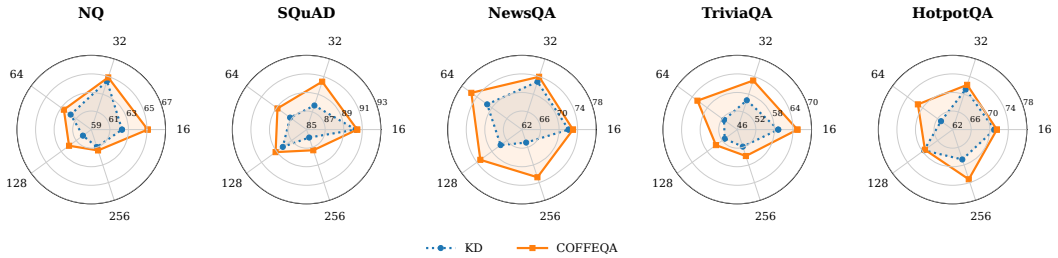


Figure 5: Teacher-student F1 agreement of standard KD and CoFFeQA across datasets and k-shot settings using T5. CoFFeQA generally achieves higher agreement than the baseline. Spokes denote k-shot settings, and radial axis F1 score ranges are adjusted independently by dataset for visualization.

Table 2: Out-of-domain F1 score comparison across k-shots for T5 and Qwen2.5 models using KD. Mean F1 scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. KD). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
DROP	KD	$38.56_{1.72}$	$35.81_{3.76}$	$\mathbf{32.48}_{2.47}$	$28.06_{3.65}$	$27.78_{1.41}$	$32.54_{4.73}$	$\mathbf{49.58}_{1.13}$	$44.92_{1.31}$	$42.07_{0.45}$	$41.28_{0.67}$	$\mathbf{42.57}_{0.94}$	$44.08_{3.36}$
	+ CoFFeQA	$\mathbf{39.24}_{1.14}$	$\mathbf{37.54}_{1.20}$	$32.41_{2.71}$	$\mathbf{32.29}_{2.00}$	$\mathbf{29.44}_{1.53}$	$\mathbf{34.19}_{4.06}$	$48.58_{0.79}$	$\mathbf{49.53}_{1.23}$	$\mathbf{45.20}_{1.31}$	$\mathbf{42.65}_{1.43}$	$41.36_{0.60}$	$\mathbf{45.46}_{3.58}$
DuoRC	KD	$50.38_{1.23}$	$48.49_{3.56}$	$\mathbf{46.98}_{2.07}$	$42.52_{1.73}$	$44.46_{1.80}$	$46.57_{3.13}$	$\mathbf{41.73}_{0.77}$	$44.84_{1.29}$	$\mathbf{43.38}_{1.30}$	$42.77_{1.16}$	$42.71_{1.49}$	$43.08_{1.15}$
	+ CoFFeQA	$\mathbf{51.66}_{0.56}$	$\mathbf{50.58}_{1.02}$	$44.55_{2.64}$	$\mathbf{46.14}_{1.23}$	$\mathbf{44.72}_{1.51}$	$\mathbf{47.53}_{3.36}$	$41.04_{1.11}$	$\mathbf{45.01}_{0.92}$	$43.11_{0.98}$	$\mathbf{43.25}_{1.00}$	$\mathbf{44.39}_{1.42}$	$\mathbf{43.36}_{1.52}$
RelationExtraction	KD	$82.96_{0.99}$	$\mathbf{83.30}_{0.46}$	$82.65_{0.61}$	$78.30_{2.66}$	$77.53_{1.47}$	$80.95_{2.79}$	$\mathbf{85.32}_{0.45}$	$85.34_{0.49}$	$84.36_{0.25}$	$82.77_{0.71}$	$82.74_{0.19}$	$84.11_{1.30}$
	+ CoFFeQA	$\mathbf{82.98}_{0.39}$	$82.99_{0.65}$	$\mathbf{83.22}_{1.22}$	$\mathbf{80.45}_{1.51}$	$\mathbf{79.55}_{1.50}$	$\mathbf{81.84}_{1.71}$	$84.71_{0.55}$	$\mathbf{85.66}_{0.33}$	$\mathbf{85.62}_{0.18}$	$\mathbf{83.61}_{0.77}$	$\mathbf{82.88}_{0.80}$	$\mathbf{84.50}_{1.23}$
AVG	KD	$57.30_{23.00}$	$55.87_{24.59}$	$\mathbf{54.04}_{25.82}$	$49.63_{25.86}$	$49.92_{25.32}$	$53.35_{24.91}$	$\mathbf{58.88}_{23.24}$	$58.37_{23.36}$	$56.60_{24.05}$	$55.61_{23.54}$	$56.00_{23.15}$	$57.09_{23.40}$
	+ CoFFeQA	$\mathbf{57.96}_{22.54}$	$\mathbf{57.03}_{23.40}$	$53.39_{26.53}$	$\mathbf{52.96}_{24.79}$	$\mathbf{51.24}_{25.69}$	$\mathbf{54.52}_{24.58}$	$58.11_{23.34}$	$\mathbf{60.07}_{22.28}$	$\mathbf{57.97}_{23.96}$	$\mathbf{56.50}_{23.48}$	$\mathbf{56.21}_{23.15}$	$\mathbf{57.77}_{23.17}$

CoFFeQA generalizes well to out-of-domain datasets. To evaluate out-of-domain generalization, we use Natural Questions for distillation and evaluate the resulting students on three out-of-domain test sets: DROP, DuoRC, and RelationExtraction. As shown in Table 2, CoFFeQA improves over standard KD when averaged across datasets and k-shot settings for both model families. For T5 and Qwen2.5, CoFFeQA improves the overall F1 score over standard KD from 53.35 to 54.52 and from 57.09 to 57.77, respectively. These results indicate that the benefits of counterfactual supervision extend beyond the in-domain datasets used for distillation. Detailed EM scores are in Appendix H.3.

CoFFeQA improves performance under modified contextual evidence. Following Longpre et al. (2021), we replace each test answer in the context with a randomly selected answer of the same entity type and evaluate against the modified target. As shown in Table 13, CoFFeQA improves average F1 agreement over standard KD from 66.72 to 69.55 for T5 and from 49.23 to 53.55 for Qwen2.5. This suggests that CoFFeQA better adapts predictions to modified contextual evidence when it differs from pre-trained knowledge. Corresponding EM results are provided in Appendix H.4.

Ablation Studies: Additionally, we conduct three ablation studies to examine different components of CoFFeQA: (i) We compare CoFFeQA with a variant that uses generated counterfactuals without CFD verification to isolate the contribution of CFD; (ii) We vary the CFD threshold τ to analyze its effect on downstream performance and teacher-student agreement; and (iii) We measure counterfactual generation costs, training time and energy consumption across baselines to analyze the computational cost of CoFFeQA. Full ablation results and experimental details are in Appendix G.

Conclusion. In this paper, we present a counterfactual-guided knowledge distillation method for few-shot open-book question answering. Motivated by the difficulty of distilling with factual samples near the teacher's decision boundary, our approach uses factual-counterfactual pairs to provide additional supervision around these challenging regions. We further introduce CFD to validate counterfactuals before distillation. Experiments across multiple datasets and LLM architectures demonstrate improved knowledge transfer under limited supervision. Limitations and future directions are discussed in Appendix B. Overall, our findings highlight the potential of counterfactual supervision for improving knowledge transfer in generative QA under limited supervision in the few-shot setting.

AI USE STATEMENT

In this work, we used generative AI tools to generate counterfactual examples as part of the proposed methodology, as described in Section 2.1. We did not use generative AI tools to generate or interpret experimental results, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for language editing, grammar correction, improving the clarity and organization of the manuscript, assisting with code development and debugging, and visualization. Counterfactual examples generated using generative AI were evaluated with the proposed CFD metric before being used for distillation. All AI-assisted text and code were reviewed and revised by the authors. We take responsibility for the final content of this work, including any text, claims, code, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the datasets, training procedures, evaluation settings, and baseline implementations in Appendix F. Specifically, Appendix F.1 describes the datasets and experimental splits, Appendix F.2 provides details of teacher model fine-tuning and evaluation, and Appendix F.3 describes the baseline formulations and implementation details. We also provide the assumptions and complete proof of our geometric analysis in Appendix E. Upon acceptance, we will publicly release the source code used to reproduce our experiments and results.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed H Chi, Jilin Chen, and Alex Beutel. Improving classifier robustness through active generative counterfactual data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 127–139, 2023.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Zero-shot llm-guided counterfactual generation: A case study on nlp model evaluation. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1243–1248. IEEE, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. Docogen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7727–7746, 2022.

Hao Chen, Rui Xia, and Jianfei Yu. Reinforced counterfactual data augmentation for dual sentiment classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 269–278, 2021.

Yudong Chen, Xuwei Xu, Frank de Hoog, Jiajun Liu, and Sen Wang. Medium-difficulty samples constitute smoothed decision boundary for knowledge distillation on pruned datasets. In *International Conference on Learning Representations*, volume 2025, pp. 54106–54123, 2025.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. Disco: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5514–5528, 2023a.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 627–638, 2023b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Pasan Dissanayake and Sanghamitra Dutta. Model reconstruction using counterfactual explanations: A perspective from polytope theory. *Advances in Neural Information Processing Systems*, 37: 83397–83429, 2024.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. Core: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2964–2984, 2022.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

David Egea, Barproda Halder, and Sanghamitra Dutta. Vision: Robust and interpretable code vulnerability detection leveraging counterfactual augmentation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 812–823, 2025.

Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59(1):17, 2026.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd workshop on machine reading for question answering*, pp. 1–13, 2019.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

Wangzhen Guo, Qinkang Gong, Yanghui Rao, and Hanjiang Lai. Counterfactual multihop qa: A cause-effect approach for reducing disconnected reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4214–4226, 2023.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust algorithmic recourse under model multiplicity with probabilistic guarantees. *IEEE Journal on Selected Areas in Information Theory*, 5:357–368, 2024.

Faisal Hamman, Pasan Dissanayake, Yanjun Fu, and Sanghamitra Dutta. Few-shot knowledge distillation of llms with counterfactual explanations. *arXiv preprint arXiv:2510.21631*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. Explaining the efficacy of counterfactually augmented data. *arXiv preprint arXiv:2010.02114*, 2020.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 163–170, 2020.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Abdullatif Köksal, Renat Aksitov, and Chung-Ching Chang. Hallucination augmented recitations for language models. *arXiv preprint arXiv:2311.07424*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pp. 333–342, 2017.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, 2024.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pp. 20852–20867. PMLR, 2023.

Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie Liu, Ge Zhang, Jiakai Wang, et al. Ddk: Distilling domain knowledge for efficient large language models. *Advances in Neural Information Processing Systems*, 37:98297–98319, 2024.

Qi Liu, Matt Kusner, and Phil Blunsom. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 187–197, 2021.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7052–7063, 2021.

Stephen McAleese and Mark Keane. A comparative analysis of counterfactual explanation methods for text classifiers. *arXiv preprint arXiv:2411.02643*, 2024.

Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*, 2021.

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, 2023.

OpenAI. GPT-4o mini: Advancing Cost-Efficient Intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024. Accessed: 2026-09-20.

Jiao Ou, Jinchao Zhang, Yang Feng, and Jie Zhou. Counterfactual data augmentation via perspective transition for open-domain dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1635–1648, 2022.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for qa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1670–1686, 2022.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Dheeraj Rajagopal, Siamak Shakeri, Cicero Nogueira dos Santos, Eduard Hovy, and Chung-Ching Chang. Counterfactual data augmentation improves factuality of abstractive summarization. *arXiv preprint arXiv:2205.12416*, 2022.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Alexis Ross, Ana Marasović, and Matthew E Peters. Explaining nlp models via minimal contrastive editing (mice). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3840–3852, 2021.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3194–3213, 2022.

Khouloud Saadi and Di Wang. Taskd-llm: Task-aware selective knowledge distillation for llms. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1876–1898, 2024.

Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1683–1693, 2018.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International symposium on intelligent data analysis*, pp. 309–318. Springer, 2001.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3645–3650, 2019.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. 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)*, pp. 4323–4332, 2019.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Ewoenam Kwaku Tokpo and Toon Calders. Fairflow: An automated approach to model-based counterfactual data augmentation for nlp. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 160–176. Springer, 2024.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, 2017.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*, 2023a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 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)*, pp. 13484–13508, 2023b.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. Cofca: A step-wise counterfactual multi-hop qa benchmark. *arXiv preprint arXiv:2402.11924*, 2024.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6707–6723, 2021.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024a.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024b.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–27, 2025.

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 *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Paul Youssef, Christin Seifert, Jörg Schlötterer, et al. Llms for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 14809–14824, 2024.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8276–8288, 2023.

A APPENDIX

CONTENTS

- **Section B** Limitations and Future Work
- **Section C** KD Performance on Boundary-Near Factual Samples
- **Section D** Counterfactual Generation Prompts
- **Section E** Proof of Proposition 1
- **Section F** Experimental Setup Details
  - **Section F.1** Datasets
  - **Section F.2** Teacher Model Fine-Tuning and Evaluation
  - **Section F.3** Additional Details of the Baselines
- **Section G** Ablation Studies
  - **Section G.1** Effect of CFD Validation
  - **Section G.2** CFD Threshold Ablation
  - **Section G.3** Cost Analysis
- **Section H** Additional Experimental Results
  - **Section H.1** In-Domain EM Results
  - **Section H.2** In-Domain Agreement Results
  - **Section H.3** Out-of-Domain EM Results
  - **Section H.4** Performance under Modified Contextual Evidence

B Limitations and Future Work

Our current counterfactual generation framework is limited to QA samples with entity-based answers. Extending counterfactual generation beyond these could broaden the applicability of CofFeQA. Future work could explore more general counterfactual generation strategies that can accommodate diverse answer types while preserving the intended contextual changes.

The current formulation of CofFeQA is also designed specifically for open-book question answering, where counterfactuals can be constructed by modifying the supporting context while preserving the question. Extending the definition, generation, and validation of counterfactuals to other generative tasks is an important direction for investigating whether counterfactual-guided few-shot distillation generalizes beyond open-book QA.

C KD Performance on Boundary-Near Factual Samples

To investigate how proximity to the teacher's decision boundary affects knowledge distillation, we group factual samples according to their prediction margins under the teacher model. Intuitively, a small margin indicates that the teacher assigns similar probabilities to its two most likely token predictions, reflecting greater uncertainty in its output. We use margin sampling (Scheffer et al., 2001) as a proxy for proximity to the teacher's decision boundary.

Measuring proximity to the decision boundary. For a factual sample $x = (Q, C, A)$ with target answer sequence $y = (y_1, \ldots, y_T)$, the teacher produces a probability distribution over the vocabulary at each answer position t. Let $v^t_{(1)}$ and $v^t_{(2)}$ denote the teacher's most likely and next most likely tokens, respectively, at position t, conditioned on the input x and the preceding target tokens $y_{<t}$. We define the token-level prediction margin as

$$m_t(x, y) = \log P_\theta\Big(v^t_{(1)} \mid x, y_{<t}\Big) - \log P_\theta\Big(v^t_{(2)} \mid x, y_{<t}\Big). \tag{4}$$

The prediction margin for the complete answer sequence is then computed as the mean of the token-level margins:

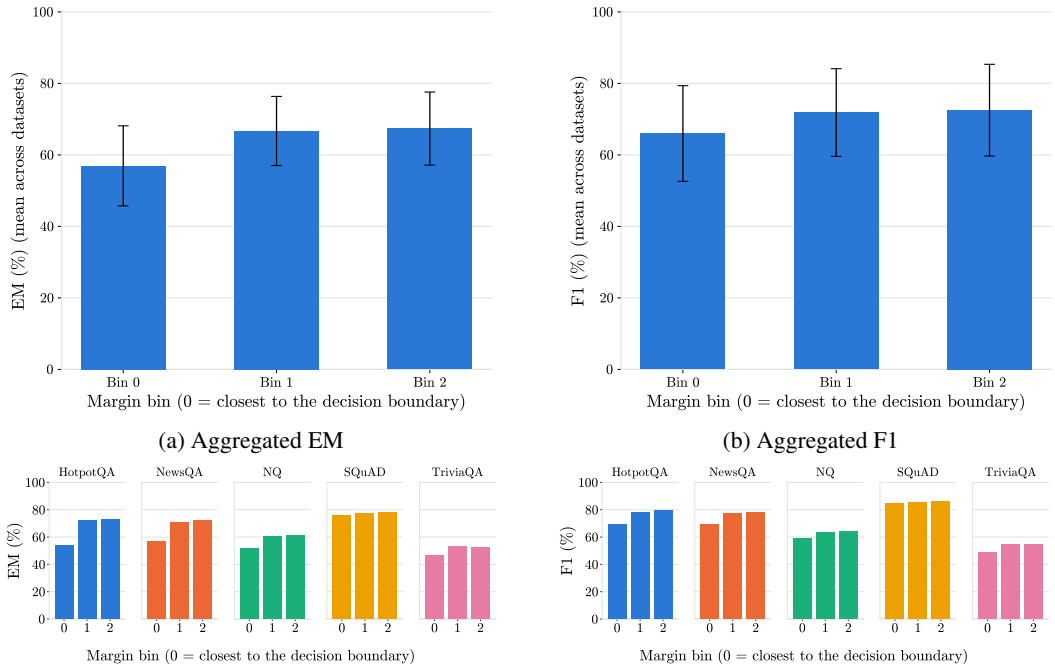
$$\text{margin}(x, y) = \frac{1}{T}\sum_{t=1}^{T} m_t(x, y). \tag{5}$$

Here, $v^t_{(1)}$ and $v^t_{(2)}$ are the teacher's own top-1 and top-2 predictions at each position and are not defined according to whether they match the ground-truth token. A smaller average margin indicates greater ambiguity between the teacher's two most likely token predictions. We therefore use this quantity as a proxy for proximity to the teacher's decision boundary.

For each dataset, we sort the factual samples in ascending order of their teacher prediction margins and divide them into three margin bins, where Bin 0 contains samples with the smallest margins, Bin 1 contains samples with intermediate margins, and Bin 2 contains samples with the largest margins. To evaluate the effect of teacher prediction margin on knowledge distillation under a controlled few-shot budget, we sample 100 factual examples from each bin and perform knowledge distillation separately using the samples from each bin. Thus, each setting uses the same distillation budget of 100 factual samples, with the primary difference being their teacher prediction margins.

Figure 6 reports the resulting KD performance when distillation is performed using 100 factual samples from each of the three margin bins. Figures 6(a) and 6(b) show EM and F1 averaged across the five in-domain datasets, respectively. For both metrics, distillation using the smallest-margin samples in Bin 0 results in substantially lower performance than using samples from Bins 1 and 2. The mean EM increases from 56.94% in Bin 0 to 67% in Bins 1–2, while the mean F1 increases from 65.99% to 72–73% in Bins 1–2. In contrast, the difference between Bins 1 and 2 is relatively small for both metrics. This suggests that the main degradation in distillation performance occurs when the few-shot factual supervision consists of samples with the smallest teacher prediction margins.

Figures 6(c) and 6(d) show the corresponding dataset-wise EM and F1 results for HotpotQA, NewsQA, Natural Questions (NQ), SQuAD, and TriviaQA. The same general pattern is observed across the five datasets: distillation using samples from Bin 0 results in lower EM and F1 than using samples from the larger-margin bins. The difference is particularly pronounced for HotpotQA, NewsQA, TriviaQA and



(a) Aggregated EM

(b) Aggregated F1

(c) Dataset-wise EM

(d) Dataset-wise F1

Figure 6: **KD performance as a function of the teacher prediction margin.** Factual samples are grouped into three bins according to their average token-level prediction margins, with Bin 0 containing the smallest margin samples and Bin 2 containing the largest margin samples. Smaller margins are used as a proxy for greater proximity to the teacher's decision boundary. Knowledge distillation is performed separately using 100 factual samples from each bin, maintaining the same distillation budget across all three settings. Panels (a) and (b) report EM and F1 averaged across the five in-domain datasets, respectively, while panels (c) and (d) report the corresponding dataset-wise results. Distillation using samples from the smallest margin bin generally results in lower performance than using samples from the larger margin bins.

NQ, whereas SQuAD exhibits a smaller difference across the three bins. Thus, the lower distillation performance associated with the smallest-margin samples is observed across datasets rather than being driven by a single benchmark.

Overall, these results support our motivating observation that factual samples with small teacher prediction margins are more challenging to distill under limited factual supervision. Since a small margin reflects greater ambiguity between the teacher's most likely token predictions, we use these samples as an operational approximation of boundary-near factual examples. Importantly, this degradation is observed while maintaining the same 100 sample distillation budget across all three margin bins. This observation motivates CoFFeQA to complement boundary-near factual samples with counterfactual supervision, providing additional information about how the teacher's response changes under modified contextual evidence.

D Counterfactual Generation Prompts

We use GPT-4o-mini (OpenAI, 2024) for counterfactual generation. The same prompts are used for both the T5 and Qwen2.5 experiments. Counterfactual generation consists of three stages: (i) identifying the entity type of the original answer, (ii) generating a plausible alternative entity of the same type, and (iii) minimally modifying the original context to support the alternative answer. The prompts used in these stages are shown in Figure 7.

Generated counterfactuals are subsequently evaluated using CFD before being used for distillation.

Entity Type Classification

What type of named entity is this answer: "{answer_text}"?
Classify as PERSON, LOCATION, ORGANIZATION, DATE, NUMERIC, or OTHER if none apply.

Alternative Entity Generation

The original answer is "{original_answer}", which is a {entity_type}.
Propose ONE plausible alternative {entity_type} that is CLOSE to "{original_answer}" (similar in scale, category, or context) but clearly DIFFERENT from it.
Do NOT propose any of these already-tried alternatives: {avoid_list}.
Respond with ONLY a JSON object:
{"alternative_entity": "<your proposed alternative>"}

Counterfactual Context Generation: System Prompt

You are an expert at creating counterfactual reading-comprehension passages for QA training data. Given an original context that supports a specific answer, you rewrite it MINIMALLY so that it instead unambiguously supports a different target answer, by replacing occurrences of the original answer (and grammatically-dependent references to it) with the target answer, while leaving everything else in the passage unchanged. Do not add explanations or extra content.

Counterfactual Context Generation: User Prompt

Question: {question}

Original context (supports answer "{original_answer}"): {original_context}

Target answer (the new counterfactual answer): "{target_answer}"

Rewrite the context above so that it minimally changes to unambiguously support "{target_answer}" as the answer to the question, instead of "{original_answer}". Replace occurrences of "{original_answer}" (and any grammatically-dependent references to it) with "{target_answer}", keeping everything else in the passage exactly the same. Do not add commentary.

Respond with ONLY a JSON object:
{"reasoning": "<1-2 sentences on what you changed>",
"counterfactual_context": "<the full rewritten context>"}

Figure 7: **Prompts used for counterfactual generation.** The generation process consists of entity type classification, alternative entity generation, and counterfactual context generation. The final stage uses a system prompt and a user prompt to minimally modify the original context to support the alternative answer.

E Proof of Proposition 1

In this section, we provide the proof of the upper bound on the Hausdorff distance between the teacher's context-response boundary $\mathcal{M}_t$ and the student's context-response boundary $\mathcal{M}_s$.

By definition, the Hausdorff distance between $\mathcal{M}_t$ and $\mathcal{M}_s$ is

$$d_H(\mathcal{M}_t, \mathcal{M}_s) = \max\left\{ \sup_{t\in\mathcal{M}_t} \inf_{s\in\mathcal{M}_s} \|t-s\|,\ \sup_{s\in\mathcal{M}_s} \inf_{t\in\mathcal{M}_t} \|s-t\| \right\}. \tag{6}$$

We bound the two directed distances separately (Hamman et al., 2025).

Part 1: Bounding the deviation from $\mathcal{M}_t$ to $\mathcal{M}_s$

Let $t \in \mathcal{M}_t$ be an arbitrary point on the teacher's boundary. By the boundary-coverage assumption (Assumption 3), there exists a teacher boundary crossing point $t_i^* \in \mathcal{M}_t$ associated with a factual-counterfactual pair (c_i, c_i') such that (see Figure 8):

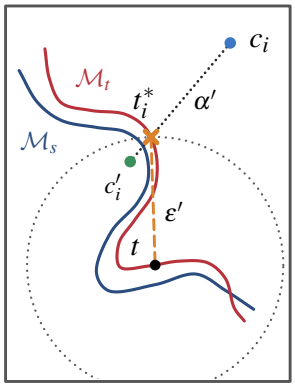


Figure 8: Visualization of the distances between $\mathcal{M}_t$ and $\mathcal{M}_s$. c_i and c_i' are factual and counterfactual pairs with $\|c_i - c_i'\| = \alpha' \leq \alpha$). t is an arbitrary point in $\mathcal{M}_t$ that finds some t_i^* cut point in $\mathcal{M}_t$ within ϵ' distance ($\epsilon' \leq \epsilon$).

$$\|t - t_i^*\| \leq \epsilon. \tag{7}$$

Under Assumption 1, the factual-counterfactual perturbation is bounded as

$$\|c_i - c_i'\| \leq \alpha. \tag{8}$$

Under the teacher-student agreement assumption (Assumption 2), the teacher and student agree on the responses associated with both endpoints c_i and c_i'. Since the factual and counterfactual contexts correspond to different responses, both the teacher and student boundaries cross the segment $[c_i, c_i']$. Let $t_i^* \in \mathcal{M}_t$ and $s_i^* \in \mathcal{M}_s$ denote their respective boundary crossing points on this segment.

Because both t_i^* and s_i^* lie on the same segment $[c_i, c_i']$,

$$\|t_i^* - s_i^*\| \leq \|c_i - c_i'\| \leq \alpha. \tag{9}$$

Combining Equations equation 7 and equation 9 and applying the triangle inequality, we obtain

$$\|t - s_i^*\| \leq \|t - t_i^*\| + \|t_i^* - s_i^*\| \leq \epsilon + \alpha. \tag{10}$$

Since there exists at least one point $s_i^* \in \mathcal{M}_s$ within distance $\epsilon + \alpha$ of t, the infimum distance satisfies:

$$\inf_{s \in \mathcal{M}_s} \|t - s\| \leq \epsilon + \alpha. \tag{11}$$

Taking the supremum over all points $t \in \mathcal{M}_t$ yields:

$$\sup_{t \in \mathcal{M}_t} \inf_{s \in \mathcal{M}_s} \|t - s\| \leq \epsilon + \alpha. \tag{12}$$

PART 2: BOUNDING THE DEVIATION FROM $\mathcal{M}_s$ TO $\mathcal{M}_t$

We follow a symmetric logic for the student boundary. For any point $s \in \mathcal{M}_s$, Assumption 3 guarantees the existence of a student boundary cutpoint s_i^* such that $\|s - s_i^*\| \leq \epsilon$.

From the relationship established in Part 1, the distance between this student cutpoint and the corresponding teacher cutpoint t_i^* is $\|s_i^* - t_i^*\| \leq \alpha$. Applying the triangle inequality:

$$\|s - t_i^*\| \leq \|s - s_i^*\| + \|s_i^* - t_i^*\| \leq \epsilon + \alpha \tag{13}$$

Since $t_i^* \in \mathcal{M}_t$, it follows that $\inf_{t \in \mathcal{M}_t} \|s - t\| \leq \epsilon + \alpha$. Taking the supremum over all $s \in \mathcal{M}_s$:

$$\sup_{s \in \mathcal{M}_s} \inf_{t \in \mathcal{M}_t} \|s - t\| \leq \epsilon + \alpha \tag{14}$$

FINAL RESULT

Combining Equations equation 12 and equation 14, we obtain

$$d_H(\mathcal{M}_t, \mathcal{M}_s) \leq \max\{\epsilon + \alpha, \epsilon + \alpha\} = \alpha + \epsilon. \tag{15}$$

This confirms that the student boundary remains within an $(\alpha + \epsilon)$-tube of the teacher boundary.

F EXPERIMENTAL SETUP DETAILS

F.1 DATASETS

We evaluate COFFEQA on eight open-domain question answering benchmarks, comprising five in-domain and three out-of-domain datasets. The in-domain benchmarks include Natural Questions (NQ), SQuAD, NewsQA, TriviaQA, and HotpotQA. NQ consists of real user queries collected from

Table 3: Statistics of teacher and student datasets used in our experiments. Teacher datasets contain the original training/validation splits, while student datasets contain available factual instances for distillation and evaluation.

	Teacher Dataset		Student Dataset		
Dataset	Train	Val.	Train	Val.	Test
NQ	93,663	10,408	34,383	3,821	2,283
SQuAD	77,847	8,651	16,341	1,816	2,790
NewsQA	66,655	7,407	10,956	1,218	690
TriviaQA	55,519	6,170	4,415	491	607
HotpotQA	65,630	7,294	28,956	3,218	2,052

Google search logs (Kwiatkowski et al., 2019). SQuAD (Stanford Question Answering Dataset) contains questions authored by crowdworkers based on Wikipedia articles (Rajpurkar et al., 2016). NewsQA is constructed from CNN news articles with questions generated crowdworkers (Trischler et al., 2017). TriviaQA consists QA pairs sourced from trivia and quiz-league websites (Joshi et al., 2017). HotpotQA focuses on multi-hop reasoning, requiring models to combine information from multiple Wikipedia articles (Yang et al., 2018). The out-of-domain benchmarks include DROP, DuoRC, and RelationExtraction. DROP contains quantitative QA pairs which has numerical and text answers (Dua et al., 2019). DuoRC contains QA pairs derived from movie plots, constructed by crowdworkers using IMDb and Wikipedia (Saha et al., 2018). RelationExtraction reformulates relation extraction as a reading comprehension task using Wikipedia text (Levy et al., 2017).

The datasets are designed for generalization in open-book question answering, where samples are drawn from diverse distributions. The datasets have following properties in common.

- **Question distribution:** The test sets contain different styles of questions, such as entity-based and relational questions, collected from diverse sources, including crowdworkers, domain experts, and exam writers.
- **Context distribution:** Passages are drawn from heterogeneous sources such as science articles, novels, medical abstracts, news articles.
- **Answerability:** The answer can be generated using the provided context, and the answer span appears within the first 800 tokens of the passage.

The training and validation sets for the teacher are drawn from the original training splits after removing duplicate entries. Dataset statistics are reported in Table 3. We generate counterfactuals only for the samples with entity-based answers. Statistics of the knowledge distillation datasets are also shown in Table 3, where student *train* and *val* denote available factual training samples with entity-based answers for distillation. The in-domain test sets for knowledge distillation are taken from the original development splits after filtering for entity-based answer samples. The out-of-domain test set counts are as follows: DROP (496), and DuoRC (450), RelationExtraction (912).

F.2 Teacher Model Fine-Tuning and Evaluation

All fine-tuning experiments were conducted on a machine equipped with four NVIDIA RTX 6000 Ada GPUs. For T5-Large fine-tuning experiments, we used a batch size of 4, a learning rate of 1×10^{-4}, and a maximum of 5 training epochs. Training was stopped early based on validation loss using an early stopping criterion with a patience of 3 validation checks. For Qwen2.5-7B-IT fine-tuning experiments, we used a batch size of 1, a learning rate of 2×10^{-5}, and a maximum of 5 training epochs. Training was stopped early based on validation loss using an early stopping criterion with a patience of 3 validation checks. On average, all models converged within 1 to 2 epochs.

We evaluated the fine-tuned teacher models using the same test sets for each dataset. The results can be found in Table 4

Table 4: Teacher model performance (EM and F1) on factual and counterfactual test sets across five in-domain QA benchmarks for T5 and Qwen2.5 models.

Test Set Type	Model	Score	NQ	SQuAD	NewsQA	TriviaQA	HotpotQA
Factual	T5	EM	74.90	79.64	73.19	67.38	77.19
		F1	77.76	87.78	78.97	69.27	85.49
	Qwen2.5	EM	79.46	83.91	75.65	73.48	79.34
		F1	82.83	91.04	82.50	76.37	88.52
Counterfactual	T5	EM	67.94	79.75	66.38	58.32	65.84
		F1	70.79	87.04	71.35	60.58	71.37
	Qwen2.5	EM	69.78	79.35	66.23	31.47	64.13
		F1	73.61	86.46	70.55	32.22	70.74

F.3 Additional details of the baselines

We compare CofFeQA against three knowledge distillation baselines. To ensure a fair comparison, all methods use the same k-shot data splits, and evaluation settings.

- **Standard Knowledge Distillation (KD).** Standard knowledge distillation (KD) (Hinton et al., 2015) transfers knowledge from a larger teacher model to a smaller student model by aligning their output distributions. Let x denote the input, $y = (y_1, \ldots, y_T)$ the ground-truth output sequence, and p_s and p_t the student and teacher model distributions, respectively. The student is first trained using the standard cross-entropy objective

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log p_s(y_t \mid x, y_{<t}). \tag{16}$$

  In addition, KD encourages the student to match the teacher's token-level output distribution at each decoding step using

$$\mathcal{L}_{\mathrm{KD}} = \sum_{t=1}^{T} \mathrm{KL}\left(p_t(\cdot \mid x, y_{<t}) \,\|\, p_s(\cdot \mid x, y_{<t})\right). \tag{17}$$

  The final training objective combines the supervised and distillation losses:

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{CE}} + (1-\alpha)\mathcal{L}_{\mathrm{KD}}, \tag{18}$$

  where $\alpha \in [0, 1]$ controls the trade-off between fitting the ground-truth sequence and matching the teacher's output distribution. Thus, standard KD transfers knowledge by aligning the teacher and student token-level distributions while retaining supervision from the ground-truth sequence.
- **Layer-wise Distillation (LWD).** Layer-wise distillation (LWD) (Jiao et al., 2020) extends output-level knowledge distillation by additionally transferring information from the intermediate representations of the teacher. Representations from selected teacher layers are aligned with the corresponding representations of the student through an auxiliary layer-wise distillation objective, providing supervision at both the intermediate representation and output levels. Let $\mathbf{h}_t^{(l_t)}$ and $\mathbf{h}_s^{(l_s)}$ denote the hidden representations from a paired teacher layer l_t and student layer l_s, respectively. Since the teacher and student may have different hidden dimensions, a learnable projection W_{l_s} is applied to the student representation before alignment. The layer-wise distillation loss is defined as

$$\mathcal{L}_{\mathrm{LWD}} = \frac{1}{|\mathcal{P}|} \sum_{(l_s, l_t) \in \mathcal{P}} \left\| W_{l_s} \mathbf{h}_s^{(l_s)} - \mathbf{h}_t^{(l_t)} \right\|_2^2, \tag{19}$$

  where $\mathcal{P}$ denotes the set of paired student and teacher layers. The final objective combines the standard KD objective with the layer-wise representation loss:

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{CE}} + (1-\alpha)\mathcal{L}_{\mathrm{KD}} + \beta \mathcal{L}_{\mathrm{LWD}}, \tag{20}$$

  where β controls the contribution of intermediate representation alignment. Thus, unlike standard KD, which transfers knowledge through the output distributions, LWD additionally encourages the student to reproduce the teacher's intermediate representations.

- **Sequence-Level Knowledge Distillation (SeqKD).** Unlike standard KD, which aligns the teacher and student token-level distributions at each decoding step, sequence-level knowledge distillation (seqKD) transfers knowledge through complete sequences generated by the teacher Kim & Rush (2016). Let x denote the input and $y = (y_1, \ldots, y_T)$ denote the output sequence, and let p_s and p_t denote the student and teacher model distributions, respectively. The sequence-level distillation objective is defined as

$$\mathcal{L}_{\text{SeqKD}} = -\sum_{y \in \mathcal{Y}} p_t(y \mid x) \log p_s(y \mid x), \tag{21}$$

where $\mathcal{Y}$ denotes the set of all possible output sequences. Since directly optimizing over all possible output sequences is intractable, seqKD approximates the teacher's sequence-level distribution using its highest-probability output sequence. Specifically, the teacher-generated sequence is obtained as

$$\hat{y} \approx \arg\max_{y \in \mathcal{Y}} p_t(y \mid x), \tag{22}$$

where beam search is used to approximate the highest probability sequence under the teacher distribution. The resulting sequence-level distillation objective is then

$$\mathcal{L}_{\text{SeqKD}} \approx -\log p_s(\hat{y} \mid x) = -\sum_{t=1}^{|\hat{y}|} \log p_s\left(\hat{y}_t \mid x, \hat{y}_{<t}\right). \tag{23}$$

Thus, while standard KD transfers knowledge by matching the teacher's token-level probability distributions, seqKD uses the complete sequence generated by the teacher as the training target for the student.

For standard KD, the student was trained using a weighted combination of the MLE loss and the token-level distillation loss. The temperature was set to 1, and the α was set to 0.5.

For LWD, student decoder layers were uniformly mapped to corresponding teacher decoder layers across the model depth. Layer-wise knowledge transfer was performed by aligning the decoder self-attention distributions of the mapped teacher and student layers using KL divergence. Attention distributions were averaged across attention heads before computing the layer-wise loss. The layer-wise distillation loss was added to the standard KD objective with a weight of 5.0.

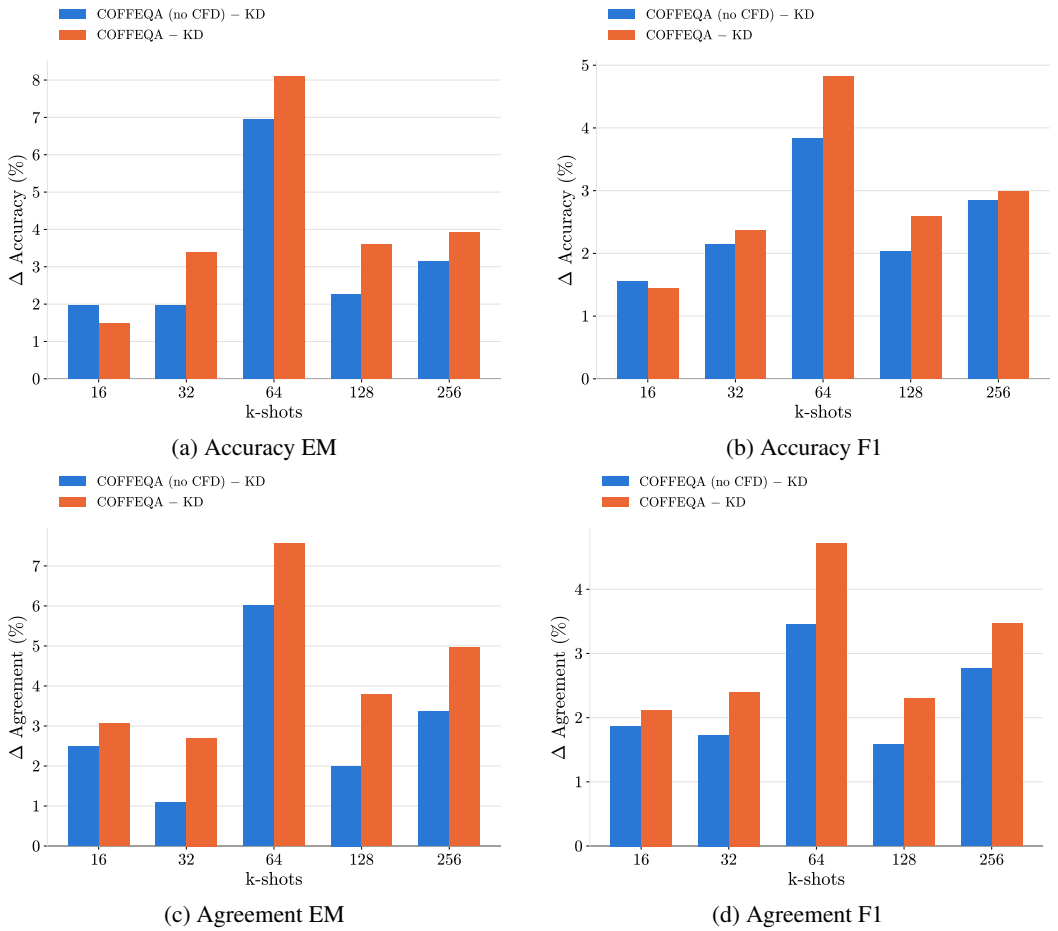
For seqKD, the teacher generated a complete output sequence for each training input using beam search with a beam size of 4. The original ground-truth answer was replaced by the corresponding teacher-generated sequence, which was then used as the training target for the student. Following the original seqKD formulation (Kim & Rush, 2016), the student was trained using cross-entropy loss on the teacher-generated sequences without an additional token-level distillation loss.

For all three baselines, we used a batch size of 4 and a learning rate of 1×10^{-4}. The maximum number of training epochs was set to 100, with early stopping based on validation loss using a patience of 3 validation checks.

G Ablation Study Results

G.1 Effect of CFD Validation

To isolate the contribution of CFD verification, we compare CoFFeQA with a variant that generates counterfactuals without CFD verification. Both variants use identical factual data splits and training settings, with the only difference being whether the generated counterfactuals are verified using CFD. For CoFFeQA with CFD verification, we use the selected threshold $\tau = 0.3$. Experiments are conducted using T5 across all datasets with five random seeds for each k-shot setting, and the results are averaged across datasets. Figure 9 reports the improvement of each variant over standard KD. CFD verification generally provides larger gains in both downstream QA performance and teacher-student agreement, with particularly clear improvements at $k = 32$, 64, and 128. These results indicate that the improvements of CoFFeQA are not solely due to the inclusion of counterfactual examples; verifying the generated counterfactuals using CFD provides additional benefits for knowledge transfer.



(a) Accuracy EM

(b) Accuracy F1

(c) Agreement EM

(d) Agreement F1

Figure 9: **Effect of CFD verification on CoFFeQA.** Performance gains over standard KD (Δ = CoFFeQA − KD) are shown across different few-shot budgets k. We compare CoFFeQA without CFD verification against the full CoFFeQA framework with CFD verification ($\tau = 0.3$). Panels (a) and (b) report accuracy EM and F1 gains, while panels (c) and (d) report teacher-student agreement EM and F1 gains. CFD verification generally yields larger improvements, demonstrating the benefit of validating generated counterfactuals before distillation.

G.2 CFD Threshold Ablation

We analyze the sensitivity of CoFFeQA to the CFD threshold τ using T5 on TriviaQA. We evaluate $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ across k, with five random seeds for each k-shot setting. For each threshold, we report the averaged improvement of CoFFeQA over standard KD (Δ = CoFFeQA − KD). As shown in Figure 10, performance is sensitive to the CFD threshold. The threshold $\tau = 0.1$ yields negligible improvements, whereas moderate thresholds produce substantially larger gains in both accuracy EM/F1 and teacher-student agreement. The gains peak around $\tau = 0.2$–0.3 and gradually decrease as the threshold becomes more permissive. Overall, these results demonstrate the importance of controlling counterfactual quality through the CFD threshold.

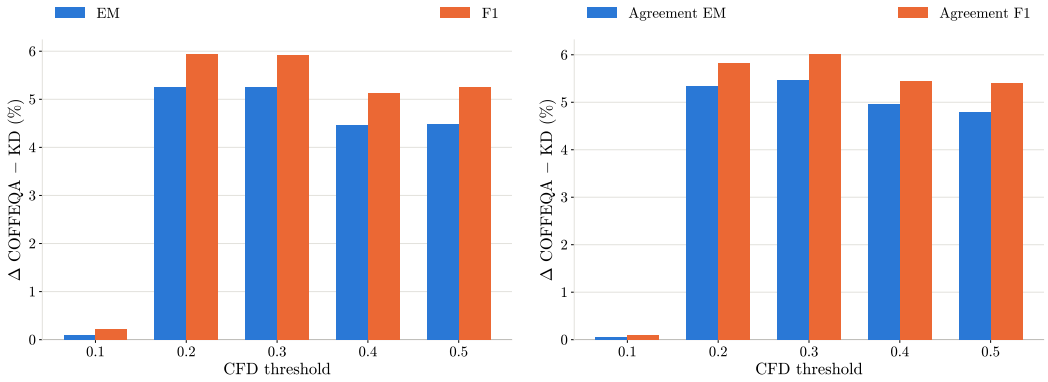


Figure 10: **CFD threshold ablation.** Effect of the CFD threshold τ on CoFFeQA using T5 on TriviaQA. The bars show the improvement of CoFFeQA over standard KD (Δ = CoFFeQA − KD), averaged across $k \in \{16, 32, 64, 128, 256\}$ and five random seeds. The left plot shows the improvement in accuracy measured by EM and F1, while the right plot shows the improvement in teacher-student agreement measured by EM and F1.

G.3 Cost Analysis

We analyze the computational cost of CoFFeQA during both counterfactual generation and distillation. Table 5 reports the per-sample cost of counterfactual generation, averaged over all generation attempts. Counterfactual generation requires approximately 2.88 seconds per candidate, with relatively small CPU and GPU energy consumption for both model settings.

Table 6 and Table 7 reports the training duration and energy consumption during distillation under T5 and Qwen2.5, respectively, measured using CodeCarbon. In both models, for standard KD and LWD, CoFFeQA achieves higher accuracy while requiring less training time and energy across all k-shot settings. For SeqKD, CoFFeQA also improves accuracy across all settings, but incurs higher computational costs at the smaller budgets ($k = 16$ and 32). From $k = 64$ onward, its computational cost becomes comparable to the baseline. Overall, these results show that CoFFeQA introduces an additional counterfactual-generation cost, while maintaining favorable computational efficiency during the subsequent distillation stage.

Table 5: Per-sample cost of counterfactual generation averaged over every candidate attempted.

Model	Duration/candidate (s)	CPU/candidate (kWh)	GPU/candidate (kWh)
T5	2.884	0.000083	0.000055
Qwen2.5	2.877	0.000071	0.000067

Table 6: Compute and energy use of baselines vs CoFFeQA during distillation under T5.

Objective	Method	k	Accuracy (EM)	Duration (s)	CPU (kWh)	GPU (kWh)
KD	Baseline	16	61.88	62.86	0.00201	0.01342
		32	60.82	50.34	0.00153	0.00980
		64	54.83	56.15	0.00170	0.01259
		128	56.07	75.90	0.00224	0.01662
		256	55.56	128.67	0.00378	0.02777
	CoFFeQA	16	63.36	45.19	0.00143	0.01016
		32	64.20	42.58	0.00133	0.00879
		64	62.93	49.46	0.00149	0.01027
		128	59.67	68.09	0.00203	0.01428
		256	59.49	114.55	0.00336	0.02487
LWD	Baseline	16	61.75	150.44	0.00474	0.02235
		32	60.94	127.87	0.00405	0.01862
		64	54.52	137.14	0.00425	0.02128
		128	55.49	161.15	0.00488	0.02327
		256	55.88	221.42	0.00694	0.03796
	CoFFeQA	16	62.94	121.86	0.00383	0.01784
		32	64.38	109.26	0.00344	0.01799
		64	62.72	127.24	0.00389	0.01743
		128	59.94	131.83	0.00400	0.01907
		256	59.78	180.71	0.00569	0.03242
SeqKD	Baseline	16	61.77	59.14	0.00186	0.01021
		32	61.19	39.74	0.00127	0.00801
		64	61.37	36.16	0.00115	0.00774
		128	61.15	44.13	0.00136	0.00988
		256	62.06	71.46	0.00215	0.01550
	CoFFeQA	16	64.81	91.77	0.00287	0.01409
		32	64.70	53.12	0.00169	0.01120
		64	64.50	37.99	0.00121	0.00832
		128	64.54	47.76	0.00149	0.00999
		256	64.06	82.79	0.00248	0.01780

Table 7: Compute and energy use of baselines vs CofFeQA during distillation under Qwen2.5.

Objective	Method	k	Accuracy (EM)	Duration (s)	CPU (kWh)	GPU (kWh)
KD	Baseline	16	61.22	748.41	0.02378	0.11453
		32	60.24	636.82	0.01930	0.11645
		64	59.13	774.16	0.02278	0.15832
		128	58.51	1202.28	0.03660	0.27748
		256	59.07	1793.38	0.05119	0.35929
	CofFeQA	16	61.73	704.12	0.02244	0.11386
		32	61.59	698.58	0.02124	0.13367
		64	61.78	693.77	0.02043	0.14167
		128	60.93	1066.53	0.03231	0.24643
		256	60.59	1728.09	0.04949	0.33793
LWD	Baseline	16	60.77	892.88	0.02964	0.14427
		32	60.94	614.12	0.01998	0.11107
		64	59.71	560.93	0.01830	0.10488
		128	59.52	641.79	0.02027	0.11110
		256	59.70	840.55	0.02688	0.16956
	CofFeQA	16	61.10	852.61	0.02920	0.13976
		32	61.51	727.77	0.02350	0.12773
		64	61.49	477.78	0.01546	0.08719
		128	61.80	526.55	0.01641	0.09615
		256	62.03	759.89	0.02396	0.14949
SeqKD	Baseline	16	59.72	584.15	0.01826	0.11342
		32	60.46	559.89	0.01701	0.10684
		64	60.96	642.25	0.01894	0.14049
		128	60.79	970.14	0.02825	0.21411
		256	62.23	1683.50	0.04842	0.38788
	CofFeQA	16	59.98	648.94	0.01986	0.12576
		32	60.72	557.76	0.01663	0.11214
		64	62.16	641.84	0.01882	0.13954
		128	62.21	925.38	0.02703	0.18591
		256	62.88	1527.88	0.04422	0.34185

H Additional Experimental Results

H.1 In-Domain EM Results

Table 8 reports the in-domain EM results across all the baselines.

Table 8: In-Domain EM score comparison across k-shots for T5 and Qwen2.5 models using baselines. Mean EM scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. baseline). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$58.10_{0.44}$	$57.73_{3.52}$	$54.73_{3.79}$	$47.84_{6.31}$	$50.32_{3.61}$	$53.75_{4.54}$	$\mathbf{61.95}_{0.36}$	$61.24_{0.61}$	$60.89_{0.44}$	$60.18_{0.53}$	$62.07_{0.47}$	$61.27_{0.78}$
	+ CoFFeQA	$\mathbf{58.89}_{0.33}$	$\mathbf{58.49}_{1.50}$	$\mathbf{56.92}_{2.34}$	$\mathbf{49.86}_{6.34}$	$\mathbf{51.20}_{5.30}$	$\mathbf{55.07}_{4.24}$	$61.79_{0.27}$	$\mathbf{63.12}_{0.35}$	$\mathbf{63.06}_{0.47}$	$\mathbf{62.66}_{0.82}$	$\mathbf{62.84}_{1.32}$	$\mathbf{62.69}_{0.54}$
	LWD	$58.20_{0.48}$	$\mathbf{57.47}_{3.89}$	$54.80_{3.04}$	$50.15_{3.73}$	$50.90_{5.43}$	$54.30_{3.68}$	$\mathbf{60.19}_{6.36}$	$61.91_{0.79}$	$61.10_{1.33}$	$60.71_{0.44}$	$61.82_{1.03}$	$61.15_{0.73}$
	+ CoFFeQA	$\mathbf{58.84}_{0.27}$	$57.45_{3.65}$	$\mathbf{55.00}_{4.16}$	$\mathbf{52.58}_{4.70}$	$\mathbf{51.16}_{6.81}$	$\mathbf{55.01}_{3.22}$	$58.95_{6.05}$	$\mathbf{62.53}_{0.63}$	$\mathbf{63.49}_{0.56}$	$\mathbf{63.48}_{0.69}$	$\mathbf{63.39}_{1.46}$	$\mathbf{62.37}_{1.95}$
	SeqKD	$54.75_{1.75}$	$54.47_{1.07}$	$56.44_{1.11}$	$55.85_{1.57}$	$\mathbf{58.59}_{1.65}$	$56.02_{1.64}$	$59.85_{0.41}$	$\mathbf{61.60}_{0.61}$	$61.96_{0.40}$	$62.37_{0.58}$	$62.94_{1.21}$	$61.74_{1.17}$
	+ CoFFeQA	$\mathbf{56.97}_{0.92}$	$\mathbf{57.66}_{1.14}$	$\mathbf{57.78}_{1.20}$	$\mathbf{58.74}_{0.76}$	$58.29_{1.49}$	$\mathbf{57.89}_{0.67}$	$\mathbf{61.72}_{0.49}$	$61.23_{0.77}$	$\mathbf{63.98}_{0.80}$	$\mathbf{63.72}_{0.76}$	$\mathbf{64.06}_{0.83}$	$\mathbf{62.94}_{1.36}$
SQuAD	KD	$78.36_{0.77}$	$75.46_{2.06}$	$73.68_{3.48}$	$74.97_{1.75}$	$74.37_{1.95}$	$75.37_{1.80}$	$\mathbf{70.27}_{0.18}$	$69.40_{0.28}$	$69.00_{0.40}$	$66.88_{0.58}$	$\mathbf{67.91}_{0.25}$	$68.69_{1.32}$
	+ CoFFeQA	$\mathbf{78.47}_{0.26}$	$\mathbf{77.86}_{1.15}$	$\mathbf{77.23}_{1.18}$	$\mathbf{77.11}_{0.46}$	$\mathbf{75.84}_{1.68}$	$\mathbf{77.30}_{0.99}$	$69.39_{0.57}$	$\mathbf{70.13}_{0.69}$	$\mathbf{69.93}_{0.55}$	$\mathbf{68.92}_{0.92}$	$67.21_{0.67}$	$\mathbf{69.12}_{1.17}$
	LWD	$78.36_{0.68}$	$75.43_{1.58}$	$74.06_{3.53}$	$73.84_{2.17}$	$74.44_{1.54}$	$75.23_{1.85}$	$\mathbf{69.40}_{0.82}$	$69.51_{0.30}$	$68.95_{0.31}$	$67.55_{0.90}$	$68.92_{1.15}$	$68.87_{0.78}$
	+ CoFFeQA	$\mathbf{78.37}_{0.33}$	$\mathbf{77.83}_{1.14}$	$\mathbf{76.54}_{1.06}$	$\mathbf{77.23}_{0.60}$	$\mathbf{76.13}_{1.79}$	$\mathbf{77.22}_{0.92}$	$68.95_{0.17}$	$\mathbf{69.85}_{0.24}$	$\mathbf{70.07}_{0.75}$	$\mathbf{69.38}_{0.31}$	$\mathbf{69.74}_{1.37}$	$\mathbf{69.60}_{0.44}$
	SeqKD	$76.45_{0.95}$	$\mathbf{77.94}_{0.93}$	$76.87_{0.67}$	$76.95_{0.18}$	$76.38_{0.64}$	$76.92_{0.62}$	$\mathbf{70.77}_{0.46}$	$70.96_{0.28}$	$70.78_{0.47}$	$70.32_{0.39}$	$71.38_{0.74}$	$70.84_{0.38}$
	+ CoFFeQA	$\mathbf{78.15}_{0.48}$	$77.30_{0.46}$	$\mathbf{77.69}_{0.71}$	$\mathbf{77.27}_{0.68}$	$\mathbf{76.56}_{0.15}$	$\mathbf{77.39}_{0.59}$	$70.04_{0.44}$	$\mathbf{71.00}_{0.43}$	$\mathbf{71.12}_{0.88}$	$\mathbf{70.78}_{0.62}$	$\mathbf{71.46}_{1.00}$	$\mathbf{70.88}_{0.53}$
NewsQA	KD	$\mathbf{64.99}_{2.59}$	$64.20_{2.93}$	$59.01_{4.29}$	$55.83_{2.73}$	$50.14_{1.27}$	$58.83_{6.15}$	$48.20_{1.05}$	$43.74_{1.28}$	$42.99_{1.22}$	$45.19_{1.65}$	$42.93_{1.48}$	$44.61_{2.21}$
	+ CoFFeQA	$64.78_{3.94}$	$\mathbf{67.59}_{1.30}$	$\mathbf{67.39}_{2.56}$	$\mathbf{64.35}_{2.37}$	$\mathbf{59.80}_{3.90}$	$\mathbf{64.78}_{3.15}$	$\mathbf{52.29}_{0.50}$	$\mathbf{49.33}_{1.11}$	$\mathbf{47.28}_{1.13}$	$\mathbf{46.67}_{1.56}$	$\mathbf{48.93}_{2.76}$	$\mathbf{48.90}_{2.20}$
	LWD	$\mathbf{64.75}_{2.54}$	$63.80_{3.28}$	$58.87_{4.49}$	$54.14_{4.04}$	$50.14_{0.74}$	$58.34_{6.24}$	$49.16_{0.55}$	$48.29_{1.11}$	$46.81_{1.47}$	$47.33_{1.21}$	$44.93_{2.77}$	$47.30_{1.60}$
	+ CoFFeQA	$64.43_{4.04}$	$\mathbf{68.06}_{0.73}$	$\mathbf{66.93}_{1.52}$	$\mathbf{63.62}_{2.63}$	$\mathbf{60.41}_{4.62}$	$\mathbf{64.69}_{3.00}$	$\mathbf{52.26}_{0.82}$	$\mathbf{49.19}_{1.21}$	$\mathbf{48.12}_{1.39}$	$\mathbf{50.23}_{0.52}$	$\mathbf{51.68}_{4.43}$	$\mathbf{50.30}_{1.71}$
	SeqKD	$54.58_{2.52}$	$55.68_{2.89}$	$58.23_{3.72}$	$60.67_{3.24}$	$62.23_{1.37}$	$58.28_{3.23}$	$41.54_{1.12}$	$43.74_{0.88}$	$45.77_{0.55}$	$48.41_{1.32}$	$52.00_{1.08}$	$46.29_{4.08}$
	+ CoFFeQA	$\mathbf{63.80}_{2.04}$	$\mathbf{66.35}_{2.23}$	$\mathbf{66.20}_{1.23}$	$\mathbf{70.09}_{1.82}$	$\mathbf{68.17}_{2.03}$	$\mathbf{66.92}_{2.36}$	$\mathbf{42.58}_{1.04}$	$\mathbf{45.71}_{1.17}$	$\mathbf{48.17}_{0.85}$	$\mathbf{50.00}_{1.99}$	$\mathbf{53.62}_{2.15}$	$\mathbf{48.02}_{4.19}$
TriviaQA	KD	$50.54_{3.88}$	$47.71_{3.13}$	$44.38_{2.08}$	$44.12_{2.82}$	$45.30_{1.17}$	$46.41_{2.71}$	$58.39_{1.08}$	$\mathbf{58.68}_{0.90}$	$56.90_{1.23}$	$54.43_{1.23}$	$\mathbf{56.05}_{0.79}$	$56.89_{1.75}$
	+ CoFFeQA	$\mathbf{56.47}_{1.45}$	$\mathbf{53.05}_{2.10}$	$\mathbf{53.15}_{6.24}$	$\mathbf{48.20}_{1.94}$	$\mathbf{47.45}_{1.73}$	$\mathbf{51.66}_{3.78}$	$\mathbf{59.04}_{0.48}$	$57.46_{0.65}$	$\mathbf{59.24}_{0.69}$	$\mathbf{57.66}_{0.92}$	$55.19_{1.49}$	$\mathbf{57.72}_{1.62}$
	LWD	$50.51_{3.54}$	$49.00_{2.51}$	$42.70_{2.28}$	$43.26_{1.75}$	$45.50_{1.07}$	$46.19_{3.46}$	$57.79_{0.32}$	$57.96_{0.83}$	$56.24_{0.30}$	$54.99_{0.59}$	$\mathbf{55.75}_{1.27}$	$56.55_{1.29}$
	+ CoFFeQA	$\mathbf{56.18}_{1.51}$	$\mathbf{52.98}_{2.33}$	$\mathbf{53.11}_{6.50}$	$\mathbf{48.01}_{1.59}$	$\mathbf{47.64}_{0.98}$	$\mathbf{51.58}_{3.66}$	$\mathbf{58.95}_{0.59}$	$\mathbf{58.42}_{0.99}$	$\mathbf{58.48}_{0.35}$	$\mathbf{57.27}_{1.17}$	$54.83_{1.27}$	$\mathbf{57.59}_{1.66}$
	SeqKD	$55.35_{3.42}$	$50.05_{1.85}$	$47.51_{2.53}$	$48.14_{1.96}$	$46.56_{1.92}$	$49.52_{3.50}$	$58.29_{0.90}$	$56.97_{0.38}$	$57.66_{0.53}$	$56.84_{1.21}$	$\mathbf{57.53}_{1.28}$	$57.46_{0.58}$
	+ CoFFeQA	$\mathbf{56.80}_{1.98}$	$\mathbf{51.86}_{1.82}$	$\mathbf{51.83}_{1.36}$	$\mathbf{48.53}_{2.97}$	$\mathbf{51.14}_{1.35}$	$\mathbf{52.03}_{2.99}$	$\mathbf{59.70}_{0.85}$	$\mathbf{59.11}_{1.01}$	$\mathbf{58.85}_{0.75}$	$\mathbf{57.23}_{0.69}$	$56.97_{0.76}$	$\mathbf{58.37}_{1.21}$
HotpotQA	KD	$57.42_{5.77}$	$58.98_{7.55}$	$42.35_{6.57}$	$57.57_{2.64}$	$57.64_{5.67}$	$54.79_{6.98}$	$\mathbf{67.31}_{0.57}$	$\mathbf{68.16}_{0.42}$	$65.86_{1.18}$	$65.88_{0.37}$	$66.40_{1.26}$	$66.72_{1.00}$
	+ CoFFeQA	$\mathbf{58.16}_{8.44}$	$\mathbf{64.03}_{4.40}$	$\mathbf{59.96}_{2.24}$	$\mathbf{58.81}_{4.43}$	$\mathbf{63.19}_{3.02}$	$\mathbf{60.83}_{2.63}$	$66.12_{1.10}$	$67.91_{0.48}$	$\mathbf{69.40}_{0.74}$	$\mathbf{68.75}_{0.75}$	$\mathbf{68.79}_{1.00}$	$\mathbf{68.19}_{1.27}$
	LWD	$\mathbf{56.93}_{5.65}$	$59.03_{8.17}$	$42.16_{6.37}$	$56.03_{5.04}$	$58.39_{5.61}$	$54.51_{7.00}$	$\mathbf{67.32}_{0.58}$	$67.03_{0.72}$	$65.46_{0.93}$	$67.03_{0.45}$	$67.10_{2.22}$	$66.79_{0.75}$
	+ CoFFeQA	$56.87_{9.31}$	$\mathbf{65.58}_{3.89}$	$\mathbf{62.01}_{1.86}$	$\mathbf{58.25}_{4.73}$	$\mathbf{63.55}_{3.92}$	$\mathbf{61.25}_{3.63}$	$66.41_{1.07}$	$\mathbf{67.55}_{0.54}$	$\mathbf{67.30}_{0.79}$	$\mathbf{68.63}_{0.97}$	$\mathbf{70.52}_{1.08}$	$\mathbf{68.08}_{1.57}$
	SeqKD	$67.69_{1.02}$	$67.82_{0.94}$	$67.81_{2.05}$	$64.14_{5.13}$	$\mathbf{66.56}_{3.21}$	$66.80_{1.58}$	$\mathbf{68.14}_{0.77}$	$\mathbf{69.02}_{0.94}$	$68.65_{0.43}$	$66.00_{0.72}$	$67.30_{1.53}$	$\mathbf{67.82}_{1.20}$
	+ CoFFeQA	$\mathbf{68.32}_{1.38}$	$\mathbf{70.32}_{0.59}$	$\mathbf{69.00}_{2.03}$	$\mathbf{68.08}_{2.52}$	$66.13_{4.35}$	$\mathbf{68.37}_{1.52}$	$65.88_{0.68}$	$66.57_{1.26}$	$\mathbf{68.68}_{0.60}$	$\mathbf{69.32}_{0.64}$	$\mathbf{68.28}_{1.59}$	$67.74_{1.46}$
AVG	KD	$61.88_{10.53}$	$60.82_{10.13}$	$54.83_{12.63}$	$56.07_{11.94}$	$55.56_{11.40}$	$57.83_{10.78}$	$61.22_{8.62}$	$60.24_{10.27}$	$59.13_{10.14}$	$58.51_{8.96}$	$59.07_{10.13}$	$59.64_{9.59}$
	+ CoFFeQA	$\mathbf{63.36}_{9.01}$	$\mathbf{64.20}_{9.42}$	$\mathbf{62.93}_{9.55}$	$\mathbf{59.67}_{11.77}$	$\mathbf{59.49}_{11.12}$	$\mathbf{61.93}_{9.98}$	$\mathbf{61.73}_{6.60}$	$\mathbf{61.59}_{8.40}$	$\mathbf{61.78}_{9.26}$	$\mathbf{60.93}_{9.25}$	$\mathbf{60.59}_{8.39}$	$\mathbf{61.32}_{8.32}$
	LWD	$61.75_{10.57}$	$60.94_{9.70}$	$54.52_{13.16}$	$55.49_{11.37}$	$55.88_{11.36}$	$57.72_{10.74}$	$60.77_{8.08}$	$60.94_{8.37}$	$59.71_{8.64}$	$59.52_{8.53}$	$59.70_{9.72}$	$60.13_{8.65}$
	+ CoFFeQA	$\mathbf{62.94}_{9.22}$	$\mathbf{64.38}_{9.67}$	$\mathbf{62.72}_{9.50}$	$\mathbf{59.94}_{11.31}$	$\mathbf{59.78}_{11.22}$	$\mathbf{61.95}_{9.96}$	$\mathbf{61.10}_{6.66}$	$\mathbf{61.51}_{8.19}$	$\mathbf{61.49}_{8.65}$	$\mathbf{61.80}_{8.08}$	$\mathbf{62.03}_{8.55}$	$\mathbf{61.59}_{7.91}$
	SeqKD	$61.77_{9.91}$	$61.19_{11.44}$	$61.37_{11.27}$	$61.15_{10.68}$	$62.06_{10.93}$	$61.51_{10.60}$	$59.72_{11.46}$	$60.46_{10.92}$	$60.96_{9.98}$	$60.79_{8.51}$	$62.23_{7.69}$	$60.83_{9.65}$
	+ CoFFeQA	$\mathbf{64.81}_{8.90}$	$\mathbf{64.70}_{10.09}$	$\mathbf{64.50}_{10.03}$	$\mathbf{64.54}_{11.13}$	$\mathbf{64.06}_{9.72}$	$\mathbf{64.52}_{9.83}$	$\mathbf{59.98}_{10.51}$	$\mathbf{60.72}_{9.59}$	$\mathbf{62.16}_{9.12}$	$\mathbf{62.21}_{8.66}$	$\mathbf{62.88}_{7.50}$	$\mathbf{61.59}_{8.95}$

H.2 In-Domain Agreement Results

Tables 9 and 10 report the in-domain teacher-student agreement results using EM and F1, respectively. For better visualization, Figures 11 and 12 illustrate the F1 agreement across datasets and k-shot settings for all baselines using T5 and Qwen2.5, respectively.

Table 9: In-domain teacher-student agreement EM score comparison across k-shots for T5 and Qwen2.5 models using baselines. Mean agreement EM scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. baseline). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$56.23_{4.87}$	$60.20_{2.87}$	$54.47_{6.56}$	$50.26_{7.34}$	$52.62_{5.00}$	$54.76_{3.76}$	$\mathbf{61.88}_{0.68}$	$62.93_{0.40}$	$62.84_{0.62}$	$62.53_{0.22}$	$63.73_{0.83}$	$62.78_{0.67}$
	+ CoFFeQA	$\mathbf{61.34}_{0.56}$	$\mathbf{60.64}_{0.49}$	$\mathbf{57.97}_{1.36}$	$\mathbf{52.94}_{6.72}$	$\mathbf{53.46}_{4.92}$	$\mathbf{57.27}_{3.93}$	$61.46_{0.53}$	$\mathbf{62.97}_{0.32}$	$\mathbf{64.52}_{0.13}$	$\mathbf{64.94}_{0.11}$	$\mathbf{64.72}_{1.17}$	$\mathbf{63.72}_{1.48}$
	LWD	$59.86_{0.63}$	$58.94_{4.18}$	$56.42_{3.20}$	$51.35_{4.46}$	$52.22_{5.87}$	$55.76_{3.85}$	$\mathbf{60.08}_{0.64}$	$63.05_{0.76}$	$62.79_{1.26}$	$62.50_{0.77}$	$62.66_{1.13}$	$62.21_{1.21}$
	+ CoFFeQA	$\mathbf{61.31}_{0.40}$	$\mathbf{58.95}_{3.37}$	$\mathbf{56.57}_{4.41}$	$\mathbf{53.98}_{4.65}$	$\mathbf{52.74}_{7.23}$	$\mathbf{56.71}_{3.52}$	$58.92_{0.62}$	$\mathbf{63.87}_{0.70}$	$\mathbf{64.87}_{0.51}$	$\mathbf{65.51}_{0.56}$	$\mathbf{65.47}_{1.02}$	$\mathbf{63.73}_{2.77}$
	SeqKD	$57.52_{1.96}$	$57.19_{1.23}$	$59.24_{1.76}$	$58.92_{1.83}$	$\mathbf{62.41}_{1.97}$	$59.06_{2.07}$	$61.45_{0.52}$	$\mathbf{64.81}_{0.70}$	$65.01_{0.38}$	$65.86_{0.44}$	$66.30_{1.21}$	$64.69_{1.91}$
	+ CoFFeQA	$\mathbf{60.22}_{0.90}$	$\mathbf{59.69}_{1.20}$	$\mathbf{60.21}_{1.00}$	$\mathbf{60.26}_{1.37}$	$60.88_{1.53}$	$\mathbf{60.25}_{0.42}$	$\mathbf{62.45}_{0.47}$	$63.09_{0.32}$	$\mathbf{66.36}_{0.48}$	$\mathbf{66.86}_{0.45}$	$\mathbf{67.15}_{0.35}$	$\mathbf{65.18}_{2.23}$
SQuAD	KD	$85.35_{0.75}$	$81.94_{1.37}$	$80.52_{3.17}$	$82.17_{2.32}$	$78.67_{6.21}$	$81.73_{2.46}$	$\mathbf{73.30}_{0.35}$	$72.70_{0.26}$	$71.99_{0.47}$	$69.98_{0.81}$	$\mathbf{71.03}_{0.69}$	$71.80_{1.32}$
	+ CoFFeQA	$\mathbf{85.73}_{0.39}$	$\mathbf{85.55}_{0.58}$	$\mathbf{83.33}_{1.33}$	$\mathbf{83.96}_{0.45}$	$\mathbf{81.28}_{2.43}$	$\mathbf{83.97}_{1.82}$	$72.18_{0.76}$	$\mathbf{73.00}_{0.50}$	$\mathbf{72.55}_{0.76}$	$\mathbf{71.63}_{0.86}$	$70.30_{0.67}$	$\mathbf{71.93}_{1.04}$
	LWD	$85.16_{1.09}$	$81.72_{2.03}$	$79.94_{4.61}$	$78.97_{2.73}$	$79.91_{2.29}$	$81.14_{2.46}$	$\mathbf{72.52}_{0.60}$	$72.51_{0.28}$	$72.19_{0.52}$	$70.53_{1.14}$	$72.28_{1.10}$	$72.01_{0.84}$
	+ CoFFeQA	$\mathbf{85.98}_{0.40}$	$\mathbf{84.37}_{1.68}$	$\mathbf{83.13}_{1.30}$	$\mathbf{84.21}_{0.77}$	$\mathbf{82.19}_{2.26}$	$\mathbf{83.97}_{1.43}$	$72.24_{0.63}$	$\mathbf{73.16}_{0.29}$	$\mathbf{73.71}_{0.73}$	$\mathbf{72.37}_{0.18}$	$\mathbf{72.94}_{1.32}$	$\mathbf{72.88}_{0.60}$
	SeqKD	$83.48_{1.46}$	$\mathbf{84.99}_{1.17}$	$83.98_{1.52}$	$\mathbf{84.57}_{0.25}$	$\mathbf{84.59}_{0.35}$	$84.32_{0.59}$	$\mathbf{74.85}_{0.45}$	$74.62_{0.53}$	$74.09_{0.78}$	$73.90_{0.72}$	$74.88_{1.16}$	$74.47_{0.45}$
	+ CoFFeQA	$\mathbf{85.56}_{0.46}$	$84.21_{0.49}$	$\mathbf{84.52}_{0.70}$	$84.47_{0.88}$	$83.25_{0.72}$	$\mathbf{84.40}_{0.82}$	$73.81_{0.65}$	$\mathbf{75.00}_{0.45}$	$\mathbf{74.65}_{0.60}$	$\mathbf{74.40}_{0.43}$	$\mathbf{75.25}_{0.70}$	$\mathbf{74.62}_{0.56}$
NewsQA	KD	$63.28_{2.18}$	$63.10_{3.40}$	$60.61_{2.80}$	$51.19_{4.13}$	$46.58_{1.56}$	$56.95_{7.62}$	$47.74_{1.38}$	$44.12_{1.68}$	$43.65_{1.56}$	$46.44_{1.30}$	$42.52_{1.92}$	$44.89_{2.14}$
	+ CoFFeQA	$\mathbf{63.83}_{2.50}$	$\mathbf{66.09}_{0.66}$	$\mathbf{67.13}_{1.61}$	$\mathbf{61.71}_{4.24}$	$\mathbf{58.00}_{3.91}$	$\mathbf{63.35}_{3.65}$	$\mathbf{51.01}_{0.64}$	$\mathbf{48.44}_{1.27}$	$\mathbf{48.17}_{1.06}$	$\mathbf{47.42}_{1.02}$	$\mathbf{50.12}_{1.82}$	$\mathbf{49.03}_{1.48}$
	LWD	$63.62_{3.09}$	$62.81_{3.88}$	$55.83_{3.91}$	$49.77_{4.38}$	$45.94_{1.27}$	$55.59_{7.81}$	$48.49_{1.44}$	$\mathbf{49.33}_{1.12}$	$48.20_{1.65}$	$47.97_{0.52}$	$45.13_{3.12}$	$47.83_{1.59}$
	+ CoFFeQA	$\mathbf{64.67}_{2.80}$	$\mathbf{66.90}_{0.42}$	$\mathbf{66.38}_{1.74}$	$\mathbf{60.75}_{3.06}$	$\mathbf{56.52}_{4.03}$	$\mathbf{63.04}_{4.37}$	$\mathbf{51.28}_{0.76}$	$48.03_{1.76}$	$\mathbf{48.61}_{1.40}$	$\mathbf{52.03}_{1.01}$	$\mathbf{52.78}_{4.17}$	$\mathbf{50.55}_{2.11}$
	SeqKD	$56.49_{3.01}$	$57.16_{2.91}$	$59.51_{3.20}$	$60.46_{2.70}$	$62.09_{1.90}$	$59.14_{2.32}$	$40.23_{1.47}$	$41.97_{1.21}$	$47.30_{0.72}$	$48.78_{1.35}$	$53.30_{1.05}$	$46.32_{5.29}$
	+ CoFFeQA	$\mathbf{64.78}_{1.38}$	$\mathbf{66.29}_{1.75}$	$\mathbf{66.29}_{1.35}$	$\mathbf{68.87}_{1.16}$	$\mathbf{66.03}_{2.30}$	$\mathbf{66.45}_{1.49}$	$\mathbf{41.28}_{1.60}$	$\mathbf{46.49}_{1.11}$	$\mathbf{48.35}_{0.85}$	$\mathbf{51.10}_{1.83}$	$\mathbf{53.94}_{1.74}$	$\mathbf{48.23}_{4.80}$
TriviaQA	KD	$56.44_{3.22}$	$53.71_{2.02}$	$49.56_{3.44}$	$49.46_{2.68}$	$50.35_{1.12}$	$51.90_{3.07}$	$\mathbf{63.26}_{1.05}$	$\mathbf{64.15}_{1.19}$	$62.57_{0.60}$	$61.25_{1.79}$	$\mathbf{63.39}_{1.28}$	$\mathbf{62.93}_{1.09}$
	+ CoFFeQA	$\mathbf{62.24}_{0.99}$	$\mathbf{59.90}_{0.68}$	$\mathbf{59.31}_{6.57}$	$\mathbf{52.62}_{2.03}$	$\mathbf{53.15}_{2.57}$	$\mathbf{57.44}_{4.31}$	$62.93_{0.78}$	$61.91_{0.41}$	$\mathbf{64.75}_{0.79}$	$\mathbf{63.76}_{1.02}$	$60.76_{2.52}$	$62.82_{1.56}$
	LWD	$55.29_{3.79}$	$54.20_{2.68}$	$47.64_{1.98}$	$48.60_{1.94}$	$49.75_{1.24}$	$51.10_{3.43}$	$62.11_{1.30}$	$\mathbf{64.38}_{0.96}$	$63.30_{1.03}$	$61.48_{0.73}$	$\mathbf{62.31}_{2.00}$	$62.72_{1.14}$
	+ CoFFeQA	$\mathbf{61.05}_{1.30}$	$\mathbf{58.42}_{2.54}$	$\mathbf{58.25}_{7.48}$	$\mathbf{52.78}_{1.70}$	$\mathbf{52.88}_{1.59}$	$\mathbf{56.68}_{3.68}$	$\mathbf{62.80}_{1.00}$	$62.67_{0.61}$	$\mathbf{64.25}_{0.69}$	$\mathbf{63.53}_{1.05}$	$61.29_{1.53}$	$\mathbf{62.91}_{1.11}$
	SeqKD	$57.83_{3.70}$	$51.83_{2.61}$	$46.92_{2.27}$	$46.85_{2.66}$	$44.48_{1.74}$	$49.58_{5.33}$	$63.30_{1.11}$	$61.78_{1.17}$	$63.26_{1.49}$	$62.47_{0.97}$	$\mathbf{65.31}_{0.86}$	$63.22_{1.32}$
	+ CoFFeQA	$\mathbf{60.46}_{1.49}$	$\mathbf{53.08}_{2.02}$	$\mathbf{52.32}_{2.40}$	$\mathbf{47.71}_{3.82}$	$\mathbf{48.90}_{1.86}$	$\mathbf{52.49}_{4.99}$	$\mathbf{63.72}_{0.50}$	$\mathbf{64.45}_{0.86}$	$\mathbf{64.55}_{0.80}$	$\mathbf{63.76}_{1.30}$	$63.06_{1.45}$	$\mathbf{63.91}_{0.61}$
HotpotQA	KD	$56.97_{4.68}$	$58.65_{7.33}$	$42.35_{7.99}$	$53.01_{4.67}$	$54.89_{6.91}$	$53.17_{6.41}$	$\mathbf{64.74}_{0.39}$	$\mathbf{66.06}_{0.32}$	$62.80_{1.31}$	$63.13_{0.77}$	$63.70_{1.31}$	$64.09_{1.33}$
	+ CoFFeQA	$\mathbf{60.36}_{4.46}$	$\mathbf{58.95}_{5.76}$	$\mathbf{57.54}_{1.95}$	$\mathbf{53.81}_{4.75}$	$\mathbf{61.99}_{3.86}$	$\mathbf{58.53}_{3.11}$	$64.05_{1.24}$	$65.77_{0.39}$	$\mathbf{66.67}_{0.79}$	$\mathbf{66.18}_{0.63}$	$\mathbf{66.34}_{0.69}$	$\mathbf{65.80}_{1.03}$
	LWD	$53.63_{5.91}$	$55.49_{8.74}$	$38.30_{6.33}$	$52.32_{5.27}$	$54.96_{5.91}$	$50.94_{7.17}$	$\mathbf{65.57}_{0.56}$	$65.70_{0.78}$	$64.05_{0.46}$	$65.05_{0.58}$	$65.60_{2.06}$	$65.19_{0.69}$
	+ CoFFeQA	$\mathbf{54.04}_{9.44}$	$\mathbf{62.83}_{4.28}$	$\mathbf{58.02}_{2.00}$	$\mathbf{54.38}_{4.91}$	$\mathbf{60.44}_{4.15}$	$\mathbf{57.94}_{3.81}$	$65.49_{1.12}$	$\mathbf{65.99}_{0.60}$	$\mathbf{66.26}_{0.61}$	$\mathbf{66.71}_{1.14}$	$\mathbf{67.86}_{0.48}$	$\mathbf{66.46}_{0.90}$
	SeqKD	$\mathbf{67.39}_{0.69}$	$66.15_{0.80}$	$66.40_{2.12}$	$61.39_{5.83}$	$\mathbf{65.40}_{3.45}$	$65.35_{2.32}$	$\mathbf{68.09}_{1.04}$	$\mathbf{68.43}_{0.41}$	$66.76_{0.83}$	$63.96_{0.46}$	$65.52_{1.40}$	$66.55_{1.85}$
	+ CoFFeQA	$65.51_{1.38}$	$\mathbf{67.13}_{1.28}$	$\mathbf{67.13}_{1.28}$	$\mathbf{65.78}_{2.88}$	$63.59_{4.39}$	$\mathbf{65.82}_{1.46}$	$67.04_{0.65}$	$67.07_{1.30}$	$\mathbf{68.36}_{0.90}$	$\mathbf{67.78}_{0.63}$	$\mathbf{66.81}_{0.96}$	$\mathbf{67.41}_{0.64}$
AVG	KD	$63.65_{12.48}$	$63.52_{10.85}$	$57.50_{14.50}$	$57.22_{14.01}$	$56.62_{12.70}$	$59.70_{12.46}$	$62.18_{9.22}$	$61.99_{10.68}$	$60.77_{10.37}$	$60.67_{8.65}$	$60.88_{10.75}$	$61.30_{9.90}$
	+ CoFFeQA	$\mathbf{66.70}_{10.72}$	$\mathbf{66.22}_{11.15}$	$\mathbf{65.06}_{10.93}$	$\mathbf{61.01}_{13.36}$	$\mathbf{61.57}_{11.60}$	$\mathbf{64.11}_{11.37}$	$\mathbf{62.33}_{7.57}$	$\mathbf{62.42}_{8.93}$	$\mathbf{63.33}_{9.07}$	$\mathbf{62.78}_{9.10}$	$\mathbf{62.45}_{7.70}$	$\mathbf{62.66}_{8.41}$
	LWD	$63.51_{12.72}$	$62.63_{11.18}$	$55.63_{15.46}$	$56.20_{12.81}$	$56.56_{13.47}$	$58.91_{12.65}$	$61.75_{8.79}$	$\mathbf{63.00}_{8.46}$	$62.10_{8.67}$	$61.51_{8.34}$	$61.59_{10.04}$	$61.99_{8.83}$
	+ CoFFeQA	$\mathbf{65.41}_{12.13}$	$\mathbf{66.29}_{10.66}$	$\mathbf{64.47}_{11.12}$	$\mathbf{61.22}_{13.22}$	$\mathbf{60.95}_{12.28}$	$\mathbf{63.67}_{11.65}$	$\mathbf{62.14}_{7.78}$	$62.75_{9.18}$	$\mathbf{63.54}_{9.17}$	$\mathbf{64.03}_{7.47}$	$\mathbf{64.07}_{7.58}$	$\mathbf{63.30}_{8.14}$
	SeqKD	$64.54_{11.47}$	$63.46_{13.09}$	$63.21_{13.57}$	$62.44_{13.69}$	$63.79_{14.25}$	$63.49_{12.94}$	$61.58_{13.01}$	$62.32_{12.34}$	$63.29_{9.84}$	$62.99_{9.09}$	$65.06_{7.69}$	$63.05_{10.31}$
	+ CoFFeQA	$\mathbf{67.31}_{10.49}$	$\mathbf{66.08}_{11.61}$	$\mathbf{66.09}_{11.88}$	$\mathbf{65.42}_{13.37}$	$\mathbf{64.53}_{12.36}$	$\mathbf{65.89}_{11.77}$	$\mathbf{61.66}_{12.22}$	$\mathbf{63.22}_{10.43}$	$\mathbf{64.45}_{9.77}$	$\mathbf{64.78}_{8.57}$	$\mathbf{65.24}_{7.72}$	$\mathbf{63.87}_{9.67}$

Table 10: In-domain teacher-student agreement F1 score comparison across k-shots for T5 and Qwen2.5 models using baselines. Mean agreement F1 scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. baseline). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$62.31_{3.15}$	$64.50_{1.70}$	$61.75_{2.79}$	$60.11_{3.06}$	$60.98_{2.11}$	$61.93_{1.66}$	$\mathbf{68.95}_{0.66}$	$69.19_{0.31}$	$68.87_{0.59}$	$68.30_{0.55}$	$68.94_{0.85}$	$68.85_{0.33}$
	+ CoFFeQA	$\mathbf{65.08}_{0.40}$	$\mathbf{64.91}_{0.42}$	$\mathbf{62.65}_{0.70}$	$\mathbf{61.96}_{2.63}$	$\mathbf{61.36}_{2.37}$	$\mathbf{63.19}_{1.71}$	$68.18_{0.32}$	$\mathbf{70.29}_{0.31}$	$\mathbf{70.47}_{0.34}$	$\mathbf{70.63}_{0.28}$	$\mathbf{70.45}_{0.76}$	$\mathbf{70.01}_{1.03}$
	LWD	$64.22_{0.62}$	$63.55_{2.69}$	$62.55_{1.45}$	$60.32_{2.31}$	$61.12_{2.68}$	$62.35_{1.63}$	$\mathbf{67.54}_{0.52}$	$69.09_{1.01}$	$69.07_{1.11}$	$68.78_{0.64}$	$69.19_{0.52}$	$68.73_{0.68}$
	+ CoFFeQA	$\mathbf{65.11}_{0.37}$	$\mathbf{63.84}_{2.50}$	$\mathbf{62.65}_{2.03}$	$\mathbf{62.31}_{1.45}$	$\mathbf{61.58}_{3.01}$	$\mathbf{63.10}_{1.39}$	$66.10_{0.51}$	$\mathbf{71.11}_{0.54}$	$\mathbf{70.93}_{0.65}$	$\mathbf{71.54}_{0.24}$	$\mathbf{71.35}_{0.52}$	$\mathbf{70.21}_{2.30}$
	SeqKD	$63.33_{1.54}$	$62.02_{1.16}$	$63.64_{1.50}$	$62.60_{1.83}$	$\mathbf{65.66}_{1.78}$	$63.45_{1.39}$	$68.48_{0.35}$	$\mathbf{71.63}_{0.55}$	$70.87_{0.41}$	$71.58_{0.11}$	$71.85_{0.78}$	$70.88_{1.39}$
	+ CoFFeQA	$\mathbf{64.85}_{0.67}$	$\mathbf{64.23}_{0.93}$	$\mathbf{64.23}_{0.78}$	$\mathbf{63.71}_{1.28}$	$64.05_{1.47}$	$\mathbf{64.21}_{0.42}$	$\mathbf{69.59}_{0.42}$	$70.19_{0.45}$	$\mathbf{72.23}_{0.57}$	$\mathbf{72.15}_{0.50}$	$\mathbf{72.30}_{0.54}$	$\mathbf{71.29}_{1.30}$
SQuAD	KD	$90.29_{0.71}$	$87.73_{0.93}$	$87.20_{1.46}$	$88.17_{1.55}$	$85.90_{3.23}$	$87.86_{1.60}$	$\mathbf{79.82}_{0.28}$	$79.14_{0.08}$	$78.58_{0.53}$	$78.14_{0.70}$	$\mathbf{78.64}_{0.43}$	$78.86_{0.64}$
	+ CoFFeQA	$\mathbf{90.45}_{0.25}$	$\mathbf{90.43}_{0.40}$	$\mathbf{88.88}_{0.85}$	$\mathbf{89.13}_{0.39}$	$\mathbf{87.34}_{1.48}$	$\mathbf{89.25}_{1.29}$	$79.14_{0.46}$	$\mathbf{79.47}_{0.57}$	$\mathbf{79.45}_{0.58}$	$\mathbf{78.70}_{0.69}$	$78.22_{0.45}$	$\mathbf{79.00}_{0.53}$
	LWD	$90.16_{0.61}$	$87.72_{1.35}$	$87.32_{2.34}$	$86.30_{1.00}$	$85.98_{1.50}$	$87.50_{1.65}$	$\mathbf{79.02}_{0.53}$	$79.18_{0.26}$	$78.92_{0.30}$	$78.33_{0.99}$	$79.56_{1.05}$	$79.00_{0.45}$
	+ CoFFeQA	$\mathbf{90.77}_{0.30}$	$\mathbf{89.78}_{0.79}$	$\mathbf{88.64}_{0.83}$	$\mathbf{89.37}_{0.57}$	$\mathbf{87.92}_{1.87}$	$\mathbf{89.30}_{1.09}$	$78.79_{0.21}$	$\mathbf{79.72}_{0.27}$	$\mathbf{80.29}_{0.39}$	$\mathbf{79.43}_{0.28}$	$\mathbf{80.14}_{1.43}$	$\mathbf{79.67}_{0.60}$
	SeqKD	$89.30_{0.97}$	$\mathbf{90.12}_{0.85}$	$\mathbf{89.54}_{0.67}$	$89.51_{0.20}$	$\mathbf{89.33}_{0.28}$	$\mathbf{89.56}_{0.33}$	$\mathbf{80.78}_{0.43}$	$80.21_{0.32}$	$80.05_{0.73}$	$79.79_{0.50}$	$80.67_{1.04}$	$80.30_{0.42}$
	+ CoFFeQA	$\mathbf{90.62}_{0.50}$	$89.30_{0.51}$	$89.43_{0.66}$	$\mathbf{89.63}_{0.45}$	$88.75_{0.71}$	$89.55_{0.68}$	$79.77_{0.44}$	$\mathbf{80.53}_{0.29}$	$\mathbf{80.45}_{0.29}$	$\mathbf{80.31}_{0.42}$	$\mathbf{81.22}_{0.64}$	$\mathbf{80.46}_{0.52}$
NewsQA	KD	$72.06_{1.88}$	$72.84_{2.12}$	$71.26_{1.39}$	$67.68_{2.77}$	$64.94_{1.45}$	$69.76_{3.34}$	$58.69_{1.01}$	$56.24_{1.74}$	$56.72_{1.78}$	$57.30_{1.50}$	$54.54_{1.80}$	$56.70_{1.52}$
	+ CoFFeQA	$\mathbf{73.00}_{2.11}$	$\mathbf{73.97}_{0.76}$	$\mathbf{75.48}_{1.49}$	$\mathbf{73.08}_{2.41}$	$\mathbf{72.80}_{1.84}$	$\mathbf{73.67}_{1.11}$	$\mathbf{61.02}_{0.67}$	$\mathbf{58.95}_{1.08}$	$\mathbf{59.33}_{0.52}$	$\mathbf{59.99}_{0.98}$	$\mathbf{61.68}_{2.02}$	$\mathbf{60.19}_{1.14}$
	LWD	$72.44_{2.55}$	$72.60_{2.24}$	$69.24_{1.92}$	$67.64_{2.49}$	$63.58_{1.11}$	$69.10_{3.74}$	$60.00_{1.41}$	$\mathbf{61.23}_{0.59}$	$\mathbf{60.02}_{1.51}$	$59.98_{1.16}$	$56.85_{2.27}$	$59.61_{1.64}$
	+ CoFFeQA	$\mathbf{73.94}_{2.16}$	$\mathbf{74.73}_{0.65}$	$\mathbf{74.81}_{1.57}$	$\mathbf{72.80}_{1.39}$	$\mathbf{71.65}_{1.83}$	$\mathbf{73.59}_{1.35}$	$\mathbf{61.54}_{0.86}$	$58.89_{1.36}$	$59.97_{1.08}$	$\mathbf{62.52}_{1.07}$	$\mathbf{63.61}_{4.34}$	$\mathbf{61.31}_{1.90}$
	SeqKD	$68.57_{1.99}$	$70.94_{1.79}$	$71.07_{2.51}$	$69.37_{1.92}$	$72.49_{1.46}$	$70.49_{1.54}$	$48.96_{1.45}$	$51.49_{0.86}$	$\mathbf{59.39}_{0.35}$	$60.59_{1.39}$	$63.64_{0.99}$	$56.81_{6.28}$
	+ CoFFeQA	$\mathbf{74.01}_{0.97}$	$\mathbf{75.27}_{1.16}$	$\mathbf{76.35}_{1.83}$	$\mathbf{76.27}_{1.39}$	$\mathbf{74.78}_{1.61}$	$\mathbf{75.33}_{1.00}$	$\mathbf{51.15}_{1.39}$	$\mathbf{57.40}_{0.32}$	$58.42_{1.31}$	$\mathbf{61.76}_{1.65}$	$\mathbf{64.65}_{1.69}$	$\mathbf{58.67}_{5.09}$
TriviaQA	KD	$59.24_{3.43}$	$55.96_{2.19}$	$51.30_{3.58}$	$51.04_{2.70}$	$51.79_{1.05}$	$53.87_{3.61}$	$65.52_{1.01}$	$\mathbf{66.50}_{1.00}$	$64.89_{0.54}$	$63.97_{1.56}$	$\mathbf{66.34}_{1.09}$	$\mathbf{65.44}_{1.05}$
	+ CoFFeQA	$\mathbf{65.44}_{0.86}$	$\mathbf{62.68}_{0.75}$	$\mathbf{61.91}_{6.87}$	$\mathbf{54.53}_{1.97}$	$\mathbf{54.98}_{2.30}$	$\mathbf{59.91}_{4.89}$	$\mathbf{65.54}_{0.55}$	$64.65_{0.42}$	$\mathbf{67.33}_{0.66}$	$\mathbf{66.12}_{0.89}$	$63.44_{2.38}$	$65.41_{1.47}$
	LWD	$58.13_{4.26}$	$56.47_{2.80}$	$49.35_{2.13}$	$50.19_{2.02}$	$51.13_{1.31}$	$53.05_{3.97}$	$64.93_{1.23}$	$\mathbf{66.71}_{0.93}$	$65.66_{0.93}$	$64.23_{0.69}$	$\mathbf{65.40}_{1.85}$	$65.39_{0.92}$
	+ CoFFeQA	$\mathbf{64.15}_{1.26}$	$\mathbf{61.39}_{2.69}$	$\mathbf{60.84}_{7.86}$	$\mathbf{54.74}_{1.74}$	$\mathbf{54.69}_{1.53}$	$\mathbf{59.16}_{4.25}$	$\mathbf{65.29}_{1.07}$	$65.22_{0.76}$	$\mathbf{66.85}_{0.73}$	$\mathbf{65.93}_{0.86}$	$64.03_{1.46}$	$\mathbf{65.47}_{1.04}$
	SeqKD	$64.78_{3.94}$	$57.82_{2.78}$	$52.95_{2.61}$	$53.12_{2.81}$	$50.10_{1.85}$	$55.75_{5.76}$	$65.54_{1.08}$	$64.28_{1.20}$	$65.49_{1.09}$	$65.37_{1.11}$	$\mathbf{68.03}_{0.88}$	$65.74_{1.38}$
	+ CoFFeQA	$\mathbf{66.54}_{1.86}$	$\mathbf{59.62}_{2.26}$	$\mathbf{58.50}_{2.34}$	$\mathbf{54.02}_{3.93}$	$\mathbf{54.05}_{1.96}$	$\mathbf{58.55}_{5.14}$	$\mathbf{66.13}_{0.46}$	$\mathbf{66.65}_{0.95}$	$\mathbf{66.84}_{0.68}$	$\mathbf{66.08}_{1.02}$	$65.89_{1.23}$	$\mathbf{66.32}_{0.41}$
HotpotQA	KD	$71.01_{2.73}$	$71.11_{4.62}$	$65.09_{3.52}$	$\mathbf{69.61}_{1.64}$	$68.77_{3.88}$	$69.12_{2.46}$	$\mathbf{74.13}_{0.47}$	$\mathbf{75.68}_{0.37}$	$72.55_{1.07}$	$73.18_{0.70}$	$73.36_{1.14}$	$73.78_{1.20}$
	+ CoFFeQA	$\mathbf{71.54}_{2.28}$	$\mathbf{72.14}_{2.83}$	$\mathbf{71.25}_{0.72}$	$69.38_{2.28}$	$\mathbf{73.26}_{2.65}$	$\mathbf{71.51}_{1.42}$	$73.66_{0.94}$	$75.04_{0.38}$	$\mathbf{76.43}_{0.66}$	$\mathbf{75.88}_{0.54}$	$\mathbf{76.24}_{1.15}$	$\mathbf{75.45}_{1.13}$
	LWD	$\mathbf{69.13}_{2.52}$	$69.97_{4.08}$	$62.94_{3.25}$	$69.41_{1.96}$	$70.02_{3.04}$	$68.29_{3.02}$	$75.34_{0.48}$	$75.91_{0.89}$	$73.95_{0.52}$	$75.46_{0.68}$	$75.56_{1.81}$	$75.24_{0.76}$
	+ CoFFeQA	$68.89_{3.68}$	$\mathbf{73.80}_{2.02}$	$\mathbf{71.17}_{1.64}$	$\mathbf{69.83}_{2.23}$	$\mathbf{73.23}_{2.24}$	$\mathbf{71.38}_{2.12}$	$\mathbf{75.37}_{1.05}$	$\mathbf{76.11}_{0.59}$	$\mathbf{76.21}_{0.50}$	$\mathbf{77.18}_{0.96}$	$\mathbf{77.99}_{0.62}$	$\mathbf{76.57}_{1.02}$
	SeqKD	$\mathbf{76.56}_{0.65}$	$76.06_{1.09}$	$75.41_{1.83}$	$73.75_{3.42}$	$\mathbf{74.70}_{3.20}$	$\mathbf{75.29}_{1.11}$	$\mathbf{76.23}_{0.70}$	$\mathbf{77.17}_{0.47}$	$75.42_{0.59}$	$72.53_{0.26}$	$74.29_{1.49}$	$75.13_{1.80}$
	+ CoFFeQA	$74.79_{1.16}$	$\mathbf{76.27}_{0.53}$	$\mathbf{76.04}_{0.88}$	$\mathbf{75.07}_{2.49}$	$73.29_{3.56}$	$75.09_{1.19}$	$75.24_{0.56}$	$75.29_{1.52}$	$\mathbf{77.08}_{0.83}$	$\mathbf{76.61}_{0.52}$	$\mathbf{75.59}_{1.35}$	$\mathbf{75.96}_{0.83}$
AVG	KD	$70.98_{12.11}$	$70.43_{11.73}$	$67.32_{13.26}$	$67.32_{13.76}$	$66.48_{12.56}$	$68.51_{12.59}$	$69.42_{8.07}$	$69.35_{8.89}$	$68.32_{8.21}$	$68.18_{8.07}$	$68.36_{9.03}$	$68.73_{8.42}$
	+ CoFFeQA	$\mathbf{73.10}_{10.33}$	$\mathbf{72.83}_{10.92}$	$\mathbf{72.03}_{11.03}$	$\mathbf{69.62}_{13.03}$	$\mathbf{69.95}_{12.44}$	$\mathbf{71.50}_{11.44}$	$\mathbf{69.51}_{7.06}$	$\mathbf{69.68}_{8.15}$	$\mathbf{70.60}_{7.91}$	$\mathbf{70.26}_{7.51}$	$\mathbf{70.01}_{7.40}$	$\mathbf{70.01}_{7.55}$
	LWD	$70.82_{12.09}$	$70.06_{11.68}$	$66.28_{13.81}$	$66.77_{13.28}$	$66.36_{12.90}$	$68.06_{12.62}$	$69.36_{7.74}$	$70.42_{7.19}$	$69.52_{7.30}$	$69.35_{7.62}$	$69.31_{8.87}$	$69.60_{7.72}$
	+ CoFFeQA	$\mathbf{72.57}_{10.88}$	$\mathbf{72.71}_{11.22}$	$\mathbf{71.62}_{11.15}$	$\mathbf{69.81}_{12.99}$	$\mathbf{69.81}_{12.64}$	$\mathbf{71.31}_{11.66}$	$\mathbf{69.42}_{7.30}$	$\mathbf{70.21}_{8.35}$	$\mathbf{70.85}_{7.94}$	$\mathbf{71.32}_{7.18}$	$\mathbf{71.42}_{7.66}$	$\mathbf{70.64}_{7.60}$
	SeqKD	$72.51_{10.70}$	$71.39_{12.69}$	$70.52_{13.62}$	$69.67_{13.55}$	$70.46_{14.28}$	$70.91_{12.77}$	$68.00_{12.24}$	$68.96_{11.49}$	$70.24_{8.12}$	$69.97_{7.33}$	$71.70_{6.44}$	$69.77_{9.01}$
	+ CoFFeQA	$\mathbf{74.16}_{10.20}$	$\mathbf{72.94}_{11.59}$	$\mathbf{72.91}_{12.02}$	$\mathbf{71.74}_{13.51}$	$\mathbf{70.98}_{12.94}$	$\mathbf{72.55}_{11.92}$	$\mathbf{68.38}_{10.96}$	$\mathbf{70.01}_{8.79}$	$\mathbf{71.00}_{8.71}$	$\mathbf{71.38}_{7.55}$	$\mathbf{71.93}_{6.88}$	$\mathbf{70.54}_{8.47}$

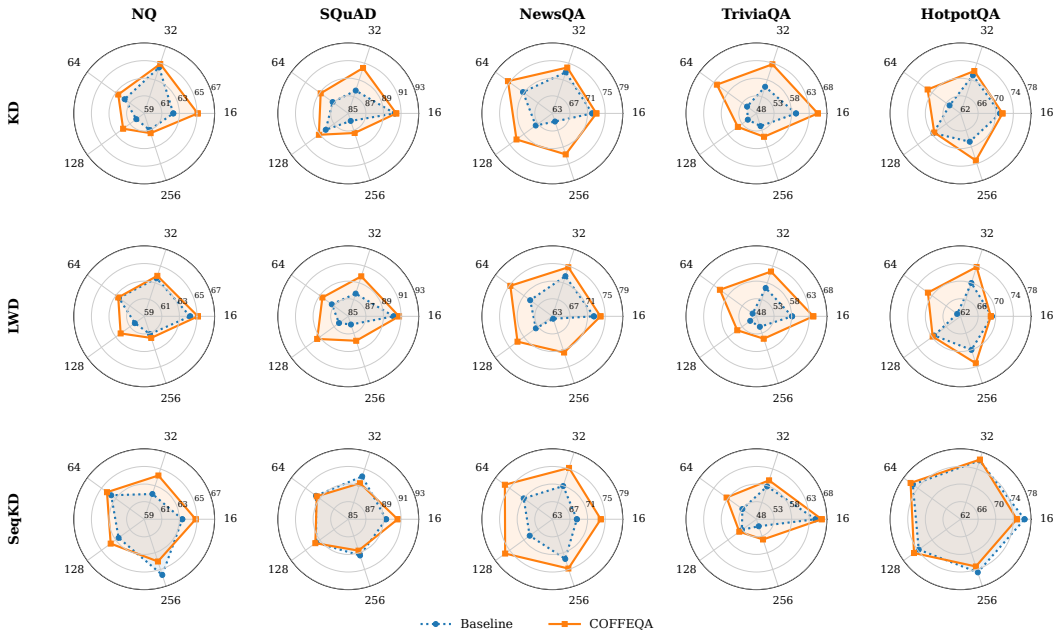


Figure 11: Teacher-student F1 agreement of KD, LWD, and SeqKD with and without CoFFeQA across datasets and k-shot settings using T5. CoFFeQA generally achieves higher agreement across the distillation baselines. Spokes denote k-shot settings, and radial axis F1 score ranges are adjusted independently by dataset for visualization.

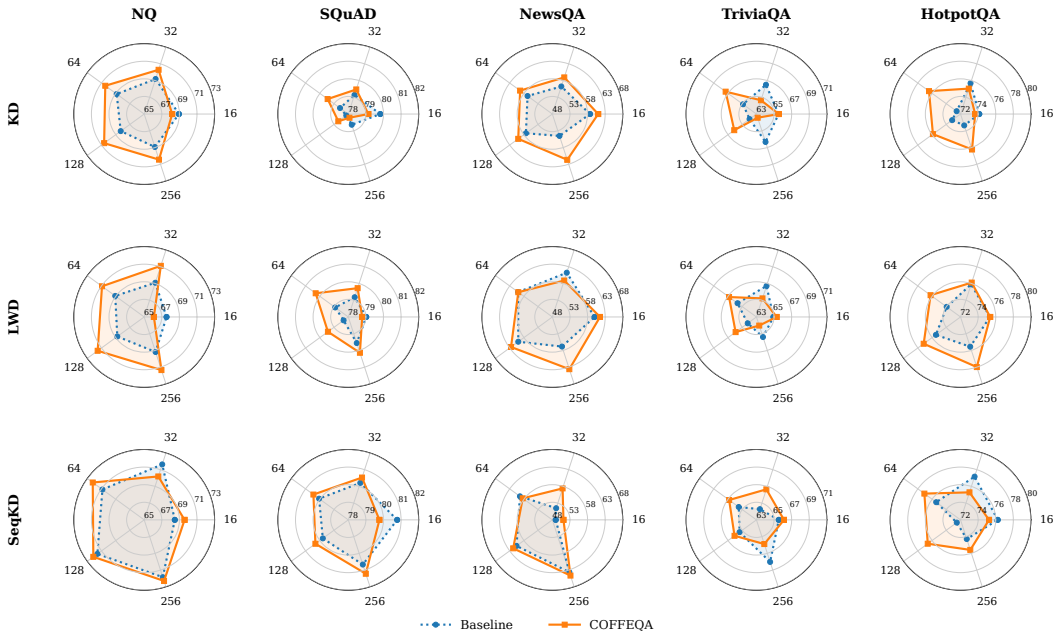


Figure 12: Teacher-student F1 agreement of KD, LWD, and SeqKD with and without CoFFeQA across datasets and k-shot settings using Qwen2.5. CoFFeQA generally achieves higher agreement across the distillation baselines. Spokes denote k-shot settings, and radial axis F1 score ranges are adjusted independently by dataset for visualization.

H.3 Out-of-Domain EM Results

Table 11 reports the out-of-domain EM results for T5 and Qwen2.5.

Table 11: Out-of-domain EM score comparison across k-shots for T5 and Qwen2.5 models using KD. Mean EM scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CofFeQA vs. KD). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
DROP	KD	$34.75_{1.78}$	$32.00_{4.64}$	$27.56_{3.50}$	$22.63_{4.29}$	$23.80_{1.97}$	$28.15_{5.20}$	$\mathbf{44.73}_{0.78}$	$39.19_{1.71}$	$36.93_{0.33}$	$35.03_{0.90}$	$\mathbf{36.57}_{1.04}$	$38.49_{3.79}$
	+ CofFeQA	$\mathbf{35.68}_{1.01}$	$\mathbf{33.86}_{1.47}$	$\mathbf{28.36}_{3.54}$	$\mathbf{25.98}_{3.56}$	$\mathbf{24.65}_{2.56}$	$\mathbf{29.71}_{4.85}$	$44.04_{0.73}$	$\mathbf{44.32}_{1.35}$	$\mathbf{39.56}_{1.51}$	$\mathbf{37.58}_{1.03}$	$35.19_{1.10}$	$\mathbf{40.14}_{4.00}$
DuoRC	KD	$44.63_{1.56}$	$43.12_{3.14}$	$\mathbf{41.47}_{1.91}$	$38.17_{1.43}$	$39.33_{1.84}$	$41.35_{2.65}$	$\mathbf{33.99}_{0.80}$	$\mathbf{38.53}_{1.06}$	$36.66_{1.60}$	$36.62_{1.10}$	$36.62_{1.80}$	$36.48_{1.62}$
	+ CofFeQA	$\mathbf{46.01}_{0.84}$	$\mathbf{44.63}_{0.87}$	$39.02_{2.37}$	$\mathbf{40.94}_{1.17}$	$\mathbf{40.04}_{1.08}$	$\mathbf{42.13}_{3.03}$	$33.68_{0.84}$	$37.55_{1.39}$	$\mathbf{36.88}_{0.78}$	$\mathbf{36.70}_{0.87}$	$\mathbf{38.58}_{1.48}$	$\mathbf{36.68}_{1.83}$
RelationExtraction	KD	$\mathbf{74.71}_{1.74}$	$\mathbf{75.13}_{0.68}$	$74.49_{1.19}$	$66.70_{5.61}$	$66.65_{3.71}$	$71.53_{4.44}$	$\mathbf{78.38}_{0.36}$	$\mathbf{78.95}_{0.41}$	$78.00_{0.17}$	$76.38_{1.13}$	$\mathbf{76.88}_{0.52}$	$77.72_{1.06}$
	+ CofFeQA	$74.62_{0.58}$	$74.34_{1.86}$	$\mathbf{75.32}_{1.82}$	$\mathbf{69.84}_{4.12}$	$\mathbf{69.22}_{2.65}$	$\mathbf{72.67}_{2.90}$	$78.29_{0.87}$	$78.73_{0.49}$	$\mathbf{79.03}_{0.38}$	$\mathbf{77.06}_{0.89}$	$75.90_{1.59}$	$\mathbf{77.80}_{1.30}$
AVG	KD	$51.36_{20.81}$	$50.08_{22.39}$	$\mathbf{47.84}_{24.11}$	$42.50_{22.35}$	$43.26_{21.70}$	$47.01_{22.24}$	$\mathbf{52.36}_{23.16}$	$52.22_{23.15}$	$50.53_{23.79}$	$49.34_{23.43}$	$\mathbf{50.02}_{23.26}$	$50.90_{23.25}$
	+ CofFeQA	$\mathbf{52.10}_{20.17}$	$\mathbf{50.94}_{20.96}$	$47.57_{24.62}$	$\mathbf{45.58}_{22.29}$	$\mathbf{44.64}_{22.64}$	$\mathbf{48.17}_{22.11}$	$52.00_{23.35}$	$\mathbf{53.53}_{22.08}$	$\mathbf{51.82}_{23.60}$	$\mathbf{50.45}_{23.05}$	$49.89_{22.59}$	$\mathbf{51.54}_{22.81}$

H.4 Performance under Modified Contextual Evidence

Tables 12 and 13 report the EM and F1 scores on modified-context test sets, respectively. Tables 14 and 15 report the corresponding teacher-student agreement scores using EM and F1, respectively.

Table 12: EM score comparison on modified-context test sets across k-shot settings for T5 and Qwen2.5 models using KD. Mean EM scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (COFFEQA vs. KD). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$54.90_{4.68}$	$58.16_{2.79}$	$53.18_{5.88}$	$49.00_{6.84}$	$51.18_{4.85}$	$53.28_{3.51}$	$\mathbf{52.10}_{0.64}$	$52.98_{0.66}$	$52.90_{0.36}$	$51.53_{0.57}$	$53.04_{0.68}$	$52.51_{0.67}$
	+ COFFEQA	$\mathbf{58.70}_{0.78}$	$\mathbf{59.33}_{0.62}$	$\mathbf{56.20}_{1.23}$	$\mathbf{51.83}_{6.47}$	$\mathbf{52.34}_{4.46}$	$\mathbf{55.68}_{3.49}$	$51.65_{0.51}$	$\mathbf{53.90}_{1.04}$	$\mathbf{56.10}_{0.50}$	$\mathbf{56.73}_{0.33}$	$\mathbf{56.93}_{1.20}$	$\mathbf{55.06}_{2.26}$
SQuAD	KD	$78.14_{0.84}$	$75.71_{0.88}$	$74.52_{2.49}$	$75.63_{1.17}$	$73.43_{5.21}$	$75.49_{1.75}$	$\mathbf{60.37}_{0.35}$	$59.74_{0.66}$	$59.18_{0.35}$	$60.34_{0.78}$	$60.98_{0.41}$	$60.12_{0.69}$
	+ COFFEQA	$\mathbf{78.42}_{0.30}$	$\mathbf{78.49}_{0.37}$	$\mathbf{77.05}_{1.20}$	$\mathbf{76.98}_{0.61}$	$\mathbf{75.77}_{1.62}$	$\mathbf{77.34}_{1.14}$	$58.48_{0.40}$	$\mathbf{60.80}_{0.38}$	$\mathbf{61.08}_{0.74}$	$\mathbf{62.57}_{1.35}$	$\mathbf{64.77}_{1.63}$	$\mathbf{61.54}_{2.32}$
NewsQA	KD	$\mathbf{64.29}_{2.04}$	$64.58_{3.03}$	$63.19_{2.52}$	$54.84_{3.97}$	$50.49_{0.94}$	$59.48_{6.43}$	$39.54_{1.29}$	$34.99_{1.04}$	$34.99_{1.38}$	$33.42_{1.22}$	$32.90_{2.06}$	$35.17_{2.61}$
	+ COFFEQA	$64.09_{3.66}$	$\mathbf{67.25}_{1.36}$	$\mathbf{67.97}_{1.77}$	$\mathbf{64.20}_{3.27}$	$\mathbf{61.91}_{3.52}$	$\mathbf{65.08}_{2.49}$	$\mathbf{46.09}_{0.98}$	$\mathbf{43.25}_{1.35}$	$\mathbf{39.42}_{1.14}$	$\mathbf{40.12}_{2.49}$	$\mathbf{40.84}_{1.85}$	$\mathbf{41.94}_{2.73}$
TriviaQA	KD	$51.60_{3.07}$	$48.43_{2.15}$	$44.68_{3.20}$	$44.38_{2.50}$	$45.50_{1.21}$	$46.92_{3.07}$	$\mathbf{38.25}_{0.63}$	$32.95_{0.74}$	$34.33_{1.89}$	$33.38_{2.00}$	$34.30_{2.85}$	$34.64_{2.11}$
	+ COFFEQA	$\mathbf{57.03}_{1.10}$	$\mathbf{54.60}_{0.90}$	$\mathbf{53.71}_{6.19}$	$\mathbf{48.11}_{1.60}$	$\mathbf{48.14}_{2.06}$	$\mathbf{52.32}_{4.02}$	$38.06_{2.03}$	$\mathbf{43.46}_{0.73}$	$\mathbf{39.51}_{1.46}$	$\mathbf{45.44}_{1.55}$	$\mathbf{44.71}_{2.20}$	$\mathbf{42.23}_{3.27}$
HotpotQA	KD	$60.63_{4.26}$	$62.12_{5.92}$	$46.31_{8.29}$	$56.83_{4.49}$	$58.26_{6.67}$	$56.83_{6.23}$	$43.39_{0.55}$	$44.14_{0.88}$	$40.87_{1.19}$	$45.36_{0.93}$	$47.45_{1.66}$	$44.24_{2.43}$
	+ COFFEQA	$\mathbf{63.17}_{4.19}$	$\mathbf{62.56}_{5.51}$	$\mathbf{61.81}_{2.05}$	$\mathbf{57.24}_{4.83}$	$\mathbf{64.81}_{3.95}$	$\mathbf{61.92}_{2.84}$	$\mathbf{45.22}_{0.40}$	$\mathbf{45.85}_{0.49}$	$\mathbf{46.25}_{0.62}$	$\mathbf{46.41}_{0.65}$	$\mathbf{54.05}_{1.30}$	$\mathbf{47.56}_{3.66}$
AVG	KD	$61.91_{10.32}$	$61.80_{9.92}$	$56.37_{12.49}$	$56.14_{11.95}$	$55.77_{10.87}$	$58.40_{10.65}$	$46.73_{9.35}$	$44.96_{11.48}$	$44.45_{11.10}$	$44.80_{11.69}$	$45.73_{12.09}$	$45.34_{11.06}$
	+ COFFEQA	$\mathbf{64.28}_{8.44}$	$\mathbf{64.44}_{9.11}$	$\mathbf{63.35}_{9.43}$	$\mathbf{59.67}_{11.42}$	$\mathbf{60.59}_{10.87}$	$\mathbf{62.47}_{9.71}$	$\mathbf{47.90}_{7.64}$	$\mathbf{49.45}_{7.68}$	$\mathbf{48.47}_{9.80}$	$\mathbf{50.25}_{9.14}$	$\mathbf{52.26}_{9.60}$	$\mathbf{49.67}_{8.51}$

Table 13: F1 score comparison on modified-context test sets across k-shot settings for T5 and Qwen2.5 models using KD. Mean F1 scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (COFFEQA vs. KD). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$60.03_{3.26}$	$61.70_{1.69}$	$59.73_{2.45}$	$57.84_{2.95}$	$58.49_{2.03}$	$59.56_{1.49}$	$\mathbf{54.52}_{0.43}$	$55.36_{0.65}$	$55.40_{0.54}$	$54.07_{0.58}$	$55.29_{0.50}$	$54.93_{0.60}$
	+ COFFEQA	$\mathbf{62.05}_{0.53}$	$\mathbf{62.65}_{0.65}$	$\mathbf{60.04}_{0.70}$	$\mathbf{59.79}_{2.52}$	$\mathbf{59.30}_{1.90}$	$\mathbf{60.76}_{1.49}$	$53.74_{0.50}$	$\mathbf{56.26}_{0.96}$	$\mathbf{58.43}_{0.45}$	$\mathbf{59.07}_{0.30}$	$\mathbf{59.51}_{1.25}$	$\mathbf{57.40}_{2.40}$
SQuAD	KD	$86.28_{0.78}$	$84.15_{0.74}$	$83.72_{1.19}$	$84.55_{1.04}$	$82.90_{2.76}$	$84.32_{1.26}$	$\mathbf{65.26}_{0.33}$	$64.49_{0.59}$	$64.19_{0.38}$	$65.67_{0.74}$	$66.22_{0.60}$	$65.17_{0.83}$
	+ COFFEQA	$\mathbf{86.52}_{0.19}$	$\mathbf{86.51}_{0.35}$	$\mathbf{85.28}_{0.98}$	$\mathbf{85.34}_{0.62}$	$\mathbf{84.31}_{1.31}$	$\mathbf{85.59}_{0.94}$	$63.82_{0.44}$	$\mathbf{65.57}_{0.28}$	$\mathbf{66.12}_{0.75}$	$\mathbf{67.46}_{1.36}$	$\mathbf{69.79}_{1.74}$	$\mathbf{66.55}_{2.23}$
NewsQA	KD	$71.99_{2.01}$	$73.56_{1.97}$	$72.39_{1.33}$	$69.03_{3.06}$	$66.17_{0.96}$	$70.63_{3.00}$	$46.19_{1.28}$	$42.10_{1.00}$	$42.91_{1.56}$	$39.78_{1.20}$	$39.19_{2.34}$	$42.03_{2.80}$
	+ COFFEQA	$\mathbf{72.81}_{3.00}$	$\mathbf{74.10}_{1.22}$	$\mathbf{75.66}_{1.73}$	$\mathbf{73.92}_{1.69}$	$\mathbf{74.23}_{1.75}$	$\mathbf{74.15}_{1.02}$	$\mathbf{52.16}_{0.74}$	$\mathbf{49.46}_{1.12}$	$\mathbf{46.00}_{1.24}$	$\mathbf{47.37}_{1.99}$	$\mathbf{47.49}_{1.25}$	$\mathbf{48.50}_{2.39}$
TriviaQA	KD	$54.79_{3.30}$	$51.31_{2.29}$	$46.77_{3.37}$	$46.18_{2.44}$	$47.35_{1.26}$	$49.28_{3.68}$	$\mathbf{40.22}_{0.69}$	$34.68_{0.70}$	$35.75_{1.90}$	$34.98_{2.06}$	$36.23_{3.00}$	$36.37_{2.24}$
	+ COFFEQA	$\mathbf{60.72}_{1.02}$	$\mathbf{57.89}_{0.89}$	$\mathbf{57.05}_{6.86}$	$\mathbf{50.47}_{1.55}$	$\mathbf{50.22}_{2.09}$	$\mathbf{55.27}_{4.70}$	$39.77_{2.14}$	$\mathbf{45.27}_{0.70}$	$\mathbf{41.43}_{1.40}$	$\mathbf{46.84}_{1.40}$	$\mathbf{46.57}_{2.29}$	$\mathbf{43.98}_{3.19}$
HotpotQA	KD	$\mathbf{71.87}_{2.37}$	$71.75_{3.96}$	$66.22_{3.20}$	$\mathbf{70.06}_{1.28}$	$69.10_{3.78}$	$69.80_{2.32}$	$47.08_{0.52}$	$47.94_{0.70}$	$44.26_{1.25}$	$48.55_{1.05}$	$50.53_{1.57}$	$47.67_{2.29}$
	+ COFFEQA	$71.73_{1.89}$	$\mathbf{72.79}_{2.70}$	$\mathbf{72.24}_{0.81}$	$69.88_{2.26}$	$\mathbf{73.22}_{2.78}$	$\mathbf{71.97}_{1.30}$	$\mathbf{49.56}_{0.43}$	$\mathbf{49.78}_{0.31}$	$\mathbf{49.92}_{0.57}$	$\mathbf{50.01}_{0.54}$	$\mathbf{57.38}_{1.43}$	$\mathbf{51.33}_{3.39}$
AVG	KD	$68.99_{12.23}$	$68.49_{12.48}$	$65.76_{13.82}$	$65.53_{14.39}$	$64.80_{13.15}$	$66.72_{13.13}$	$50.66_{9.62}$	$48.92_{11.56}$	$48.50_{11.24}$	$48.61_{12.09}$	$49.49_{12.21}$	$49.23_{11.25}$
	+ COFFEQA	$\mathbf{70.76}_{10.37}$	$\mathbf{70.79}_{11.12}$	$\mathbf{70.06}_{11.59}$	$\mathbf{67.88}_{13.36}$	$\mathbf{68.26}_{13.45}$	$\mathbf{69.55}_{11.89}$	$\mathbf{51.81}_{8.63}$	$\mathbf{53.27}_{7.92}$	$\mathbf{52.38}_{9.90}$	$\mathbf{54.15}_{8.91}$	$\mathbf{56.15}_{9.56}$	$\mathbf{53.55}_{8.75}$

Table 14: Teacher-student agreement EM score comparison on modified-context test sets across k-shot settings for T5 and Qwen2.5 models using KD. Mean agreement EM scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. KD). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$56.23_{4.87}$	$60.20_{2.87}$	$54.47_{6.56}$	$50.26_{7.34}$	$52.62_{5.00}$	$54.76_{3.76}$	$\mathbf{53.69}_{0.61}$	$55.75_{0.87}$	$55.44_{0.53}$	$54.31_{0.82}$	$55.25_{0.27}$	$54.89_{0.86}$
	+ CoFFeQA	$\mathbf{61.34}_{0.56}$	$\mathbf{60.64}_{0.49}$	$\mathbf{57.97}_{1.36}$	$\mathbf{52.94}_{6.72}$	$\mathbf{53.46}_{4.92}$	$\mathbf{57.27}_{3.93}$	$52.62_{0.58}$	$\mathbf{56.01}_{1.03}$	$\mathbf{57.34}_{0.62}$	$\mathbf{58.57}_{0.68}$	$\mathbf{58.50}_{0.62}$	$\mathbf{56.61}_{2.46}$
SQuAD	KD	$85.35_{0.75}$	$81.94_{1.37}$	$80.52_{3.17}$	$82.17_{2.32}$	$78.67_{6.21}$	$81.73_{2.46}$	$\mathbf{61.86}_{0.51}$	$61.42_{0.73}$	$61.10_{0.59}$	$61.37_{1.08}$	$61.69_{0.77}$	$61.49_{0.29}$
	+ CoFFeQA	$\mathbf{85.73}_{0.39}$	$\mathbf{85.55}_{0.58}$	$\mathbf{83.33}_{1.33}$	$\mathbf{83.96}_{0.45}$	$\mathbf{81.28}_{2.43}$	$\mathbf{83.97}_{1.82}$	$60.45_{0.43}$	$\mathbf{62.22}_{0.32}$	$\mathbf{62.22}_{1.01}$	$\mathbf{63.67}_{0.99}$	$\mathbf{64.48}_{0.86}$	$\mathbf{62.61}_{1.55}$
NewsQA	KD	$63.28_{2.18}$	$63.10_{3.40}$	$60.61_{2.80}$	$51.19_{4.13}$	$46.58_{1.56}$	$56.95_{7.62}$	$41.80_{1.25}$	$37.86_{0.82}$	$37.51_{0.50}$	$36.87_{1.29}$	$35.97_{1.75}$	$38.00_{2.24}$
	+ CoFFeQA	$\mathbf{63.83}_{2.50}$	$\mathbf{66.09}_{0.66}$	$\mathbf{67.13}_{1.61}$	$\mathbf{61.71}_{4.24}$	$\mathbf{58.00}_{3.91}$	$\mathbf{63.35}_{3.65}$	$\mathbf{46.12}_{1.78}$	$\mathbf{43.59}_{0.93}$	$\mathbf{40.84}_{0.80}$	$\mathbf{41.91}_{1.62}$	$\mathbf{43.30}_{1.75}$	$\mathbf{43.15}_{1.99}$
TriviaQA	KD	$56.44_{3.22}$	$53.71_{2.02}$	$49.56_{3.44}$	$49.46_{2.68}$	$50.35_{1.12}$	$51.90_{3.07}$	$\mathbf{39.04}_{1.14}$	$\mathbf{40.13}_{1.66}$	$38.25_{0.98}$	$\mathbf{38.16}_{0.98}$	$\mathbf{38.06}_{1.68}$	$\mathbf{38.73}_{0.88}$
	+ CoFFeQA	$\mathbf{62.24}_{0.99}$	$\mathbf{59.90}_{0.68}$	$\mathbf{59.31}_{6.57}$	$\mathbf{52.62}_{2.03}$	$\mathbf{53.15}_{2.57}$	$\mathbf{57.44}_{4.31}$	$37.46_{1.34}$	$37.69_{0.63}$	$\mathbf{39.67}_{1.35}$	$35.42_{0.26}$	$34.53_{1.08}$	$36.96_{2.03}$
HotpotQA	KD	$56.97_{4.68}$	$58.65_{7.33}$	$42.35_{7.99}$	$53.01_{4.67}$	$54.89_{6.91}$	$53.17_{6.41}$	$50.38_{0.70}$	$50.92_{0.64}$	$46.89_{1.09}$	$50.65_{1.52}$	$51.49_{1.81}$	$50.07_{1.82}$
	+ CoFFeQA	$\mathbf{60.36}_{4.46}$	$\mathbf{58.95}_{5.76}$	$\mathbf{57.54}_{1.95}$	$\mathbf{53.81}_{4.75}$	$\mathbf{61.99}_{3.86}$	$\mathbf{58.53}_{3.11}$	$\mathbf{51.86}_{0.89}$	$\mathbf{52.23}_{0.58}$	$\mathbf{52.81}_{0.67}$	$\mathbf{52.18}_{0.93}$	$\mathbf{56.61}_{1.45}$	$\mathbf{53.14}_{1.97}$
AVG	KD	$63.65_{12.48}$	$63.52_{10.85}$	$57.50_{14.50}$	$57.22_{14.01}$	$56.62_{12.70}$	$59.70_{12.46}$	$49.35_{9.21}$	$49.21_{10.08}$	$47.84_{10.41}$	$48.27_{10.56}$	$48.49_{11.12}$	$48.63_{10.22}$
	+ CoFFeQA	$\mathbf{66.70}_{10.72}$	$\mathbf{66.22}_{11.15}$	$\mathbf{65.06}_{10.93}$	$\mathbf{61.01}_{13.36}$	$\mathbf{61.57}_{11.60}$	$\mathbf{64.11}_{11.37}$	$\mathbf{49.70}_{8.54}$	$\mathbf{50.35}_{9.78}$	$\mathbf{50.57}_{10.00}$	$\mathbf{50.35}_{11.65}$	$\mathbf{51.48}_{12.23}$	$\mathbf{50.49}_{10.35}$

Table 15: Teacher-student agreement F1 score comparison on modified-context test sets across k-shot settings for T5 and Qwen2.5 models using KD. Mean agreement F1 scores are reported with standard deviations as subscripts ($Mean_{std}$). Bold values indicate the better-performing method (CoFFeQA vs. KD). Blue shading highlights aggregate results.

		T5						Qwen2.5					
Dataset	Method	16	32	64	128	256	Avg	16	32	64	128	256	Avg
NQ	KD	$62.31_{3.15}$	$64.50_{1.70}$	$61.75_{2.79}$	$60.11_{3.06}$	$60.98_{2.11}$	$61.93_{1.66}$	$\mathbf{59.30}_{0.46}$	$61.21_{0.76}$	$60.87_{0.35}$	$59.53_{0.70}$	$60.45_{0.37}$	$60.27_{0.83}$
	+ CoFFeQA	$\mathbf{65.08}_{0.40}$	$\mathbf{64.91}_{0.42}$	$\mathbf{62.65}_{0.70}$	$\mathbf{61.96}_{2.63}$	$\mathbf{61.36}_{2.37}$	$\mathbf{63.19}_{1.71}$	$58.04_{0.48}$	$\mathbf{61.55}_{1.04}$	$\mathbf{62.83}_{0.51}$	$\mathbf{63.71}_{0.58}$	$\mathbf{64.03}_{0.59}$	$\mathbf{62.03}_{2.43}$
SQuAD	KD	$90.29_{0.71}$	$87.73_{0.93}$	$87.20_{1.46}$	$88.17_{1.55}$	$85.90_{3.23}$	$87.86_{1.60}$	$\mathbf{67.69}_{0.47}$	$67.24_{0.39}$	$66.98_{0.45}$	$68.04_{0.86}$	$68.34_{0.57}$	$67.66_{0.56}$
	+ CoFFeQA	$\mathbf{90.45}_{0.25}$	$\mathbf{90.43}_{0.40}$	$\mathbf{88.88}_{0.85}$	$\mathbf{89.13}_{0.39}$	$\mathbf{87.34}_{1.48}$	$\mathbf{89.25}_{1.29}$	$66.49_{0.57}$	$\mathbf{67.98}_{0.16}$	$\mathbf{68.37}_{0.80}$	$\mathbf{69.49}_{0.87}$	$\mathbf{71.00}_{1.13}$	$\mathbf{68.66}_{1.69}$
NewsQA	KD	$72.06_{1.88}$	$72.84_{2.12}$	$71.26_{1.39}$	$67.68_{2.77}$	$64.94_{1.45}$	$69.76_{3.34}$	$50.55_{1.41}$	$47.13_{1.02}$	$47.65_{1.17}$	$45.17_{1.52}$	$44.41_{2.23}$	$46.98_{2.40}$
	+ CoFFeQA	$\mathbf{73.00}_{2.11}$	$\mathbf{73.97}_{0.76}$	$\mathbf{75.48}_{1.49}$	$\mathbf{73.08}_{2.41}$	$\mathbf{72.80}_{1.84}$	$\mathbf{73.67}_{1.11}$	$\mathbf{54.76}_{1.54}$	$\mathbf{52.54}_{1.06}$	$\mathbf{50.42}_{1.16}$	$\mathbf{51.80}_{1.19}$	$\mathbf{51.95}_{1.57}$	$\mathbf{52.29}_{1.58}$
TriviaQA	KD	$59.24_{3.43}$	$55.96_{2.19}$	$51.30_{3.58}$	$51.04_{2.70}$	$51.79_{1.05}$	$53.87_{3.61}$	$\mathbf{41.48}_{1.15}$	$\mathbf{42.62}_{1.27}$	$40.53_{0.74}$	$\mathbf{40.39}_{1.21}$	$\mathbf{40.48}_{1.58}$	$\mathbf{41.10}_{0.96}$
	+ CoFFeQA	$\mathbf{65.44}_{0.86}$	$\mathbf{62.68}_{0.75}$	$\mathbf{61.91}_{6.87}$	$\mathbf{54.53}_{1.97}$	$\mathbf{54.98}_{2.30}$	$\mathbf{59.91}_{4.89}$	$39.75_{1.33}$	$40.07_{0.55}$	$\mathbf{42.01}_{1.32}$	$37.77_{0.19}$	$37.38_{1.09}$	$39.40_{1.88}$
HotpotQA	KD	$71.01_{2.73}$	$71.11_{4.62}$	$65.09_{3.52}$	$\mathbf{69.61}_{1.64}$	$68.77_{3.88}$	$69.12_{2.46}$	$56.90_{0.71}$	$57.62_{0.58}$	$53.65_{1.08}$	$57.20_{1.38}$	$58.01_{1.66}$	$56.68_{1.74}$
	+ CoFFeQA	$\mathbf{71.54}_{2.28}$	$\mathbf{72.14}_{2.83}$	$\mathbf{71.25}_{0.72}$	$69.38_{2.28}$	$\mathbf{73.26}_{2.65}$	$\mathbf{71.51}_{1.42}$	$\mathbf{58.74}_{0.70}$	$\mathbf{59.12}_{0.62}$	$\mathbf{59.70}_{0.64}$	$\mathbf{58.93}_{0.87}$	$\mathbf{63.02}_{1.12}$	$\mathbf{59.90}_{1.78}$
AVG	KD	$70.98_{12.11}$	$70.43_{11.73}$	$67.32_{13.26}$	$67.32_{13.76}$	$66.48_{12.56}$	$68.51_{12.59}$	$55.18_{9.82}$	$55.16_{10.13}$	$53.94_{10.46}$	$54.07_{11.19}$	$54.34_{11.59}$	$54.54_{10.58}$
	+ CoFFeQA	$\mathbf{73.10}_{10.33}$	$\mathbf{72.83}_{10.92}$	$\mathbf{72.03}_{11.03}$	$\mathbf{69.62}_{13.03}$	$\mathbf{69.95}_{12.44}$	$\mathbf{71.50}_{11.44}$	$\mathbf{55.56}_{9.83}$	$\mathbf{56.25}_{10.60}$	$\mathbf{56.67}_{10.46}$	$\mathbf{56.34}_{12.24}$	$\mathbf{57.47}_{13.14}$	$\mathbf{56.46}_{11.18}$