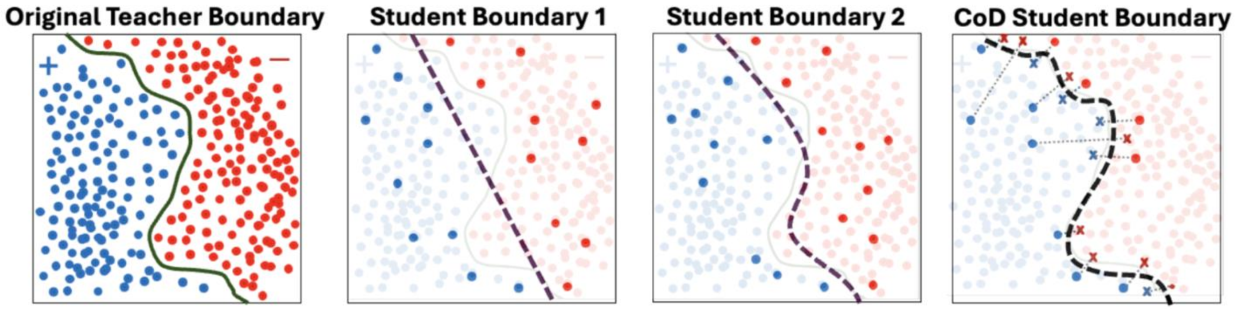
Few-Shot Knowledge Distillation of LLMs With Counterfactual Explanations

Faisal Hamman, Pasan Dissanayake, Yanjun Fu, Sanghamitra Dutta

University of Maryland, College Park



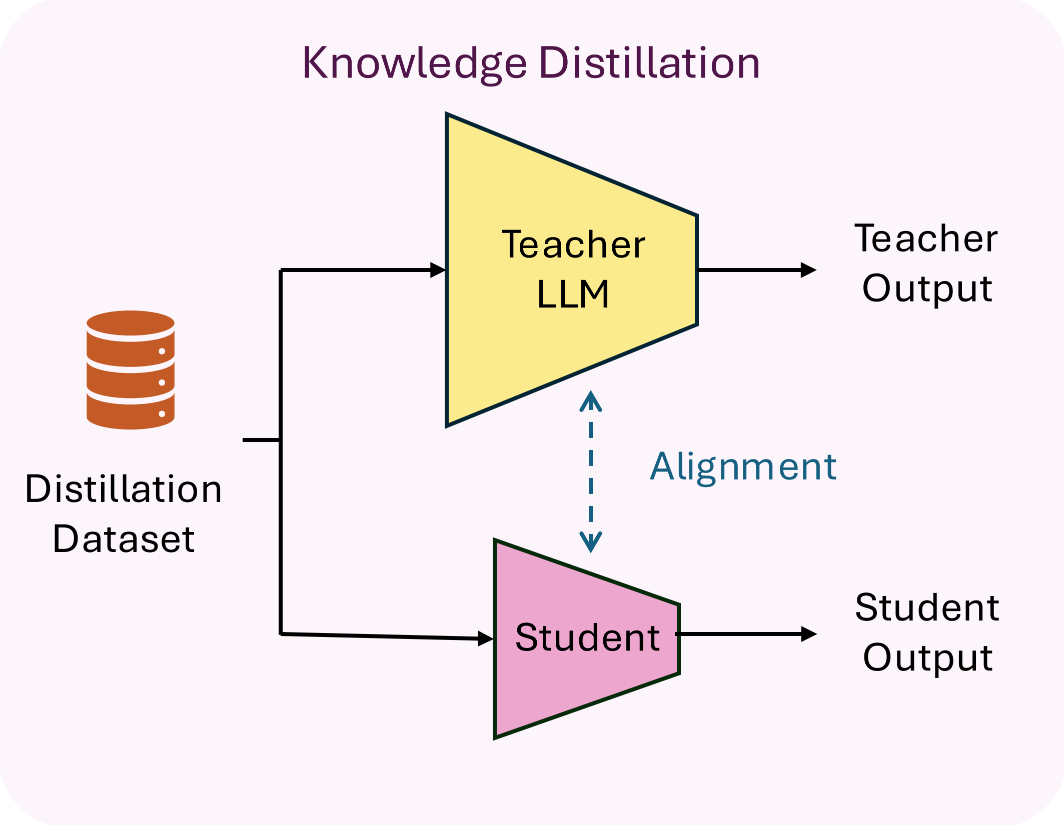
(a) Full Data (b) Few Shot Data (Set 1) (c) Few Shot Data (Set 2) (d) Half of Set 1 & CFEs

Motivations for Few-Shot Knowledge Distillation

- **LLMs** are powerful but **expensive** to deploy.

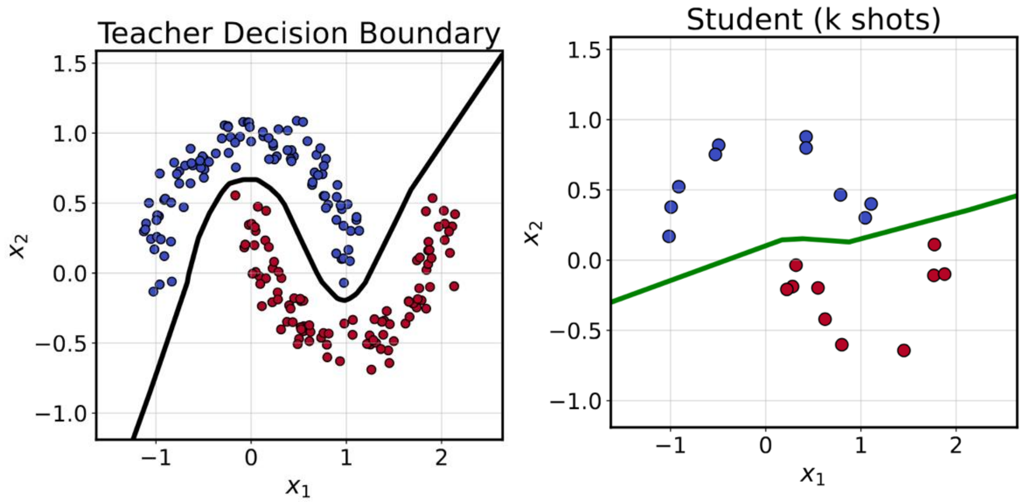
- Knowledge Distillation (KD) helps transfer capabilities to smaller models.

KD needs lots of data. For some tasks, we often have very **few** labeled samples.



Contributions

- **Few-shot** distillation leads to **poor generalization** and unfaithful student.



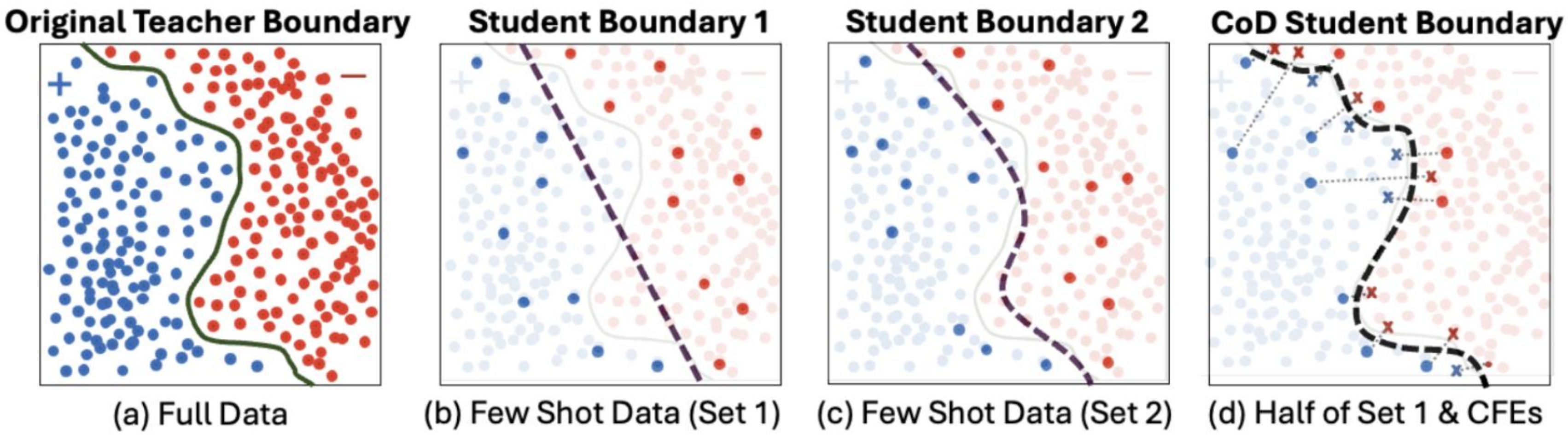
Can we use **explanations** to guide **better distillation**?

Main Contributions

1. A **counterfactual-explanation** based strategy for few-shot distillation framework.
2. **Theoretical motivations** (statistical + geometric) on why CFEs improve boundary alignment.
3. **Empirical gains** – outperform standard KD using **only half the samples.**

A counterfactual explanation-based strategy for distillation

- Leverages CFEs: minimally perturbed inputs that flip the teacher's prediction.
- CFEs lie near the decision boundary ⇒ act as "**boundary pegs**".



CFEs **align student boundary with teacher's boundary more effectively** with fewer samples!

Statistical Motivation for CFE infusion

Theorem 1 (CFEs Improve Model Parameter Estimation). *Let* $\mathbf{w}_s$ *and* $\mathbf{w}_s^{(\mathrm{cf})}$ *be the student parameters obtained via MLE on* $\mathcal{D}$ *(standard) and* $\mathcal{D}_{\mathrm{cf}}$ *(CFE-infused). Assuming the teacher's parameters* $\mathbf{w}_t$ *capture the true data-generating distribution, that CFEs lie near the decision boundary, and that the second moments* $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] \approx \mathbb{E}_{\mathbf{x}_c}[\mathbf{x}_c\mathbf{x}_c^\top]$. *Then estimation error satisfies:*

$$\mathbb{E}\left[\|\mathbf{w}_s^{(\mathrm{cf})} - \mathbf{w}_t\|^2\right] < \mathbb{E}\left[\|\mathbf{w}_s - \mathbf{w}_t\|^2\right].$$

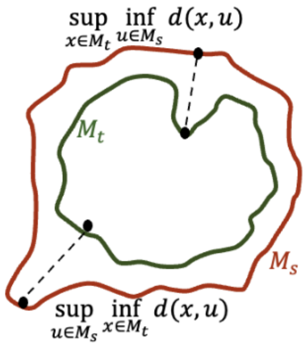
In logistic regression setting, **student's expected estimation error is lower** when training with **CFEs infused data**.

Proof relies on showing that the Fisher Information matrix is maximized with CFE infused data.

Geometric Insight for Using CFEs for Distillation

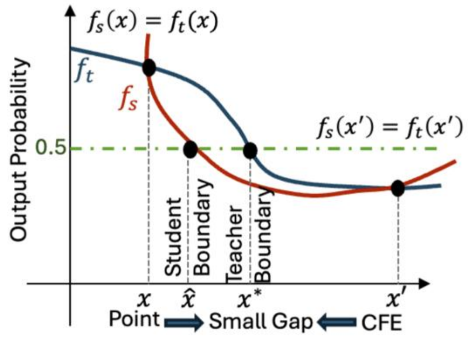
Theorem 2 (Teacher–Student Boundary Proximity). *Let $f_t,\ f_s : \mathbb{R}^{n\times d} \to; [0,1]$ be the teacher and student model, with decision boundaries $\mathcal{M}_t = \{\mathbf{x} \mid f_t(\mathbf{x}) = 0.5\}$ and $\mathcal{M}_s = \{\mathbf{x} \mid f_s(\mathbf{x}) = 0.5\}$, respectively. Assume we observe a* CFE-infused dataset $\mathcal{D}_{cf} = \{(\mathbf{x}_i, \mathbf{x}'_i)\}_{i=1}^{k}$ *satisfying: (A1) Minimal perturbation: $\|\mathbf{x}_i - \mathbf{x}'_i\|_F \leq \alpha$ with $\alpha > 0$; (A2) Exact distillation: $f_s(\mathbf{x}_i) = f_t(\mathbf{x}_i)$ and $f_s(\mathbf{x}'_i) = f_t(\mathbf{x}'_i)$; and (A3) ε-spread along the teacher and student boundary, i.e., for each pair, there exist a teacher's (or student's) crossing point $\mathbf{x}^\star_i = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{x}'_i$ for $\alpha \in (0,1)$ such that $f_t(x^\star_i) = 0.5$ (or, $f_s(x^\star_i) = 0.5$) and for every $a \in \mathcal{M}_t$ (or $\mathcal{M}_s$), there exists an i with $\|a - \mathbf{x}^\star_i\|_2 \leq \varepsilon$. Then the Hausdorff distance between the decision boundaries obeys:*

$$\mathrm{H}(\mathcal{M}_s, \mathcal{M}_t) \leq \alpha + \varepsilon.$$

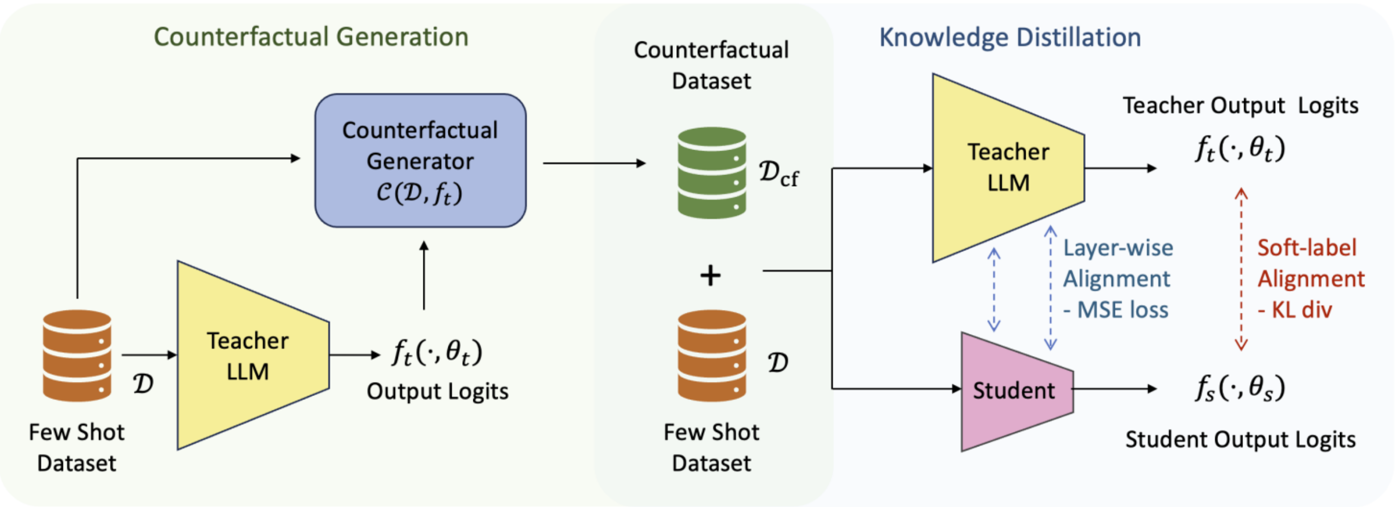


Hausdorff Distance

CFEs act as boundary anchors that **pull the student's decision surface toward the teacher's**, ensuring their boundaries stay within a tight $(\alpha + \varepsilon)$-tube and **yielding faithful few-shot distillation**.



Proposed Algorithm: **Co**unterfactual-Explanation-infused-**D**istillation (**CoD**)



Algorithm 1 CoD: CFE-infused Distillation

Require: Teacher g_t, student g_s, dataset $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{k}$, CFGen, learning rate η, loss weights α (KD), β (LWD), Epochs E

```
D_cf ← ∅
for all (x, y) ∈ D_k do
    x' ← CFGen(x, g_t)
    D_cf ← D_cf ∪ {(x', 1 − y)}
end for
D_train ← D_k ∪ D_cf
for e = 1 to E do
    for all (x, y) ∈ D_train do
        L_hard ← CE(g_s(x), y)
        L_KD ← KL(g_t(x) || g_s(x))
        L_LWD ← Σ_{l∈I} ||h_t^(l) − h_s^(l)||_2^2
        L ← L_hard + α L_KD + β L_LWD
        Update θ_s ← θ_s − η ∇_{θ_s} L
    end for
end for
return distilled student g_s
```

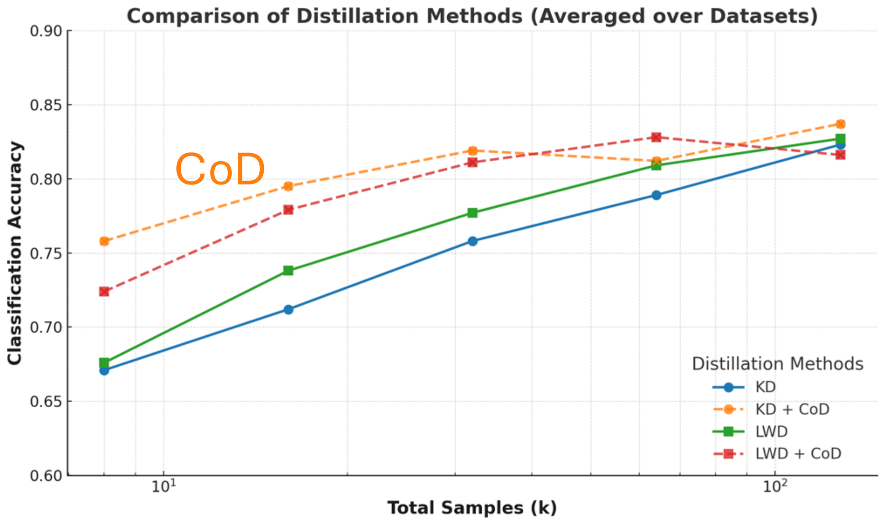
Sentiment Analysis Classification Task

Original Input
*I **liked** the movie.*

Counterfactual
*I **disliked** the movie.*

Empirical Validations

- Generated CFEs using LLMs (GPT4o)
- Baselines: KD, LWD, TED
- 6 NLP datasets
- Models: DeBERTa-v3, Qwen-2.5

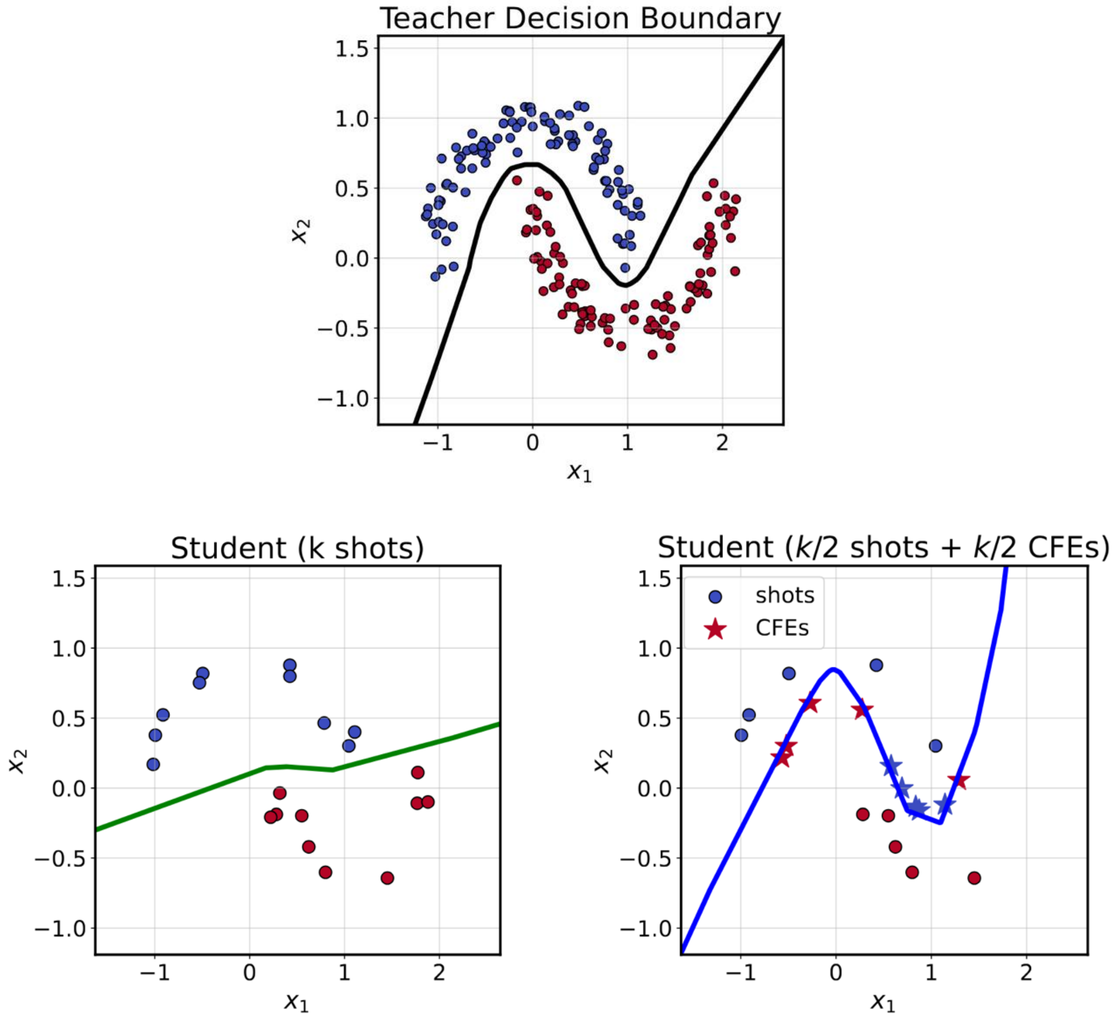


Dataset	Method	Total Samples (k) 8	16	32	64	128	512
Amazon Polarity	KD	0.671 ±0.046	0.712 ±0.033	0.758 ±0.032	0.789 ±0.022	0.823 ±0.016	0.846 ±0.007
	+CoD	**0.758** ±0.027	**0.795** ±0.033	**0.819** ±0.035	**0.812** ±0.004	**0.837** ±0.014	**0.860** ±0.015
	LWD	0.676 ±0.090	0.738 ±0.033	0.777 ±0.009	0.809 ±0.015	**0.827** ±0.025	**0.842** ±0.019
	+CoD	**0.724** ±0.052	**0.779** ±0.056	**0.811** ±0.015	**0.828** ±0.015	0.816 ±0.020	0.841 ±0.013
CoLA	KD	0.693 ±0.062	0.707 ±0.029	0.721 ±0.012	0.747 ±0.005	0.758 ±0.009	0.771 ±0.003
	+CoD	**0.739** ±0.026	**0.755** ±0.017	**0.769** ±0.011	**0.769** ±0.016	**0.772** ±0.006	**0.791** ±0.004
	LWD	0.713 ±0.031	0.698 ±0.037	0.731 ±0.021	0.744 ±0.007	0.750 ±0.018	0.761 ±0.011
	+ CoD	**0.730** ±0.035	**0.744** ±0.031	**0.762** ±0.011	**0.752** ±0.009	**0.756** ±0.010	**0.784** ±0.003
IMDB	KD	0.714 ±0.047	0.817 ±0.028	0.875 ±0.027	0.896 ±0.008	**0.912** ±0.009	**0.917** ±0.006
	+ CoD	**0.835** ±0.078	**0.888** ±0.005	**0.890** ±0.011	**0.899** ±0.007	0.907 ±0.006	0.913 ±0.005
	LWD	0.760 ±0.046	0.836 ±0.045	0.875 ±0.024	0.889 ±0.013	0.905 ±0.008	**0.914** ±0.006
	+ CoD	**0.861** ±0.017	**0.886** ±0.011	**0.893** ±0.006	**0.898** ±0.005	0.905 ±0.010	0.913 ±0.010
SST2	KD	0.617 ±0.042	0.712 ±0.052	0.757 ±0.063	0.820 ±0.019	0.848 ±0.013	**0.899** ±0.007
	+ CoD	**0.719** ±0.063	**0.781** ±0.034	**0.821** ±0.013	**0.827** ±0.008	**0.853** ±0.015	0.892 ±0.018
	LWD	0.627 ±0.053	0.721 ±0.055	0.776 ±0.031	0.817 ±0.005	0.829 ±0.013	**0.892** ±0.012
	+ CoD	**0.694** ±0.079	**0.785** ±0.028	**0.832** ±0.011	**0.830** ±0.007	**0.835** ±0.012	0.880 ±0.020
Yelp	KD	0.714 ±0.058	0.817 ±0.031	0.855 ±0.021	**0.878** ±0.006	0.885 ±0.018	**0.916** ±0.007
	+ CoD	**0.740** ±0.094	**0.832** ±0.045	**0.860** ±0.018	0.874 ±0.006	**0.888** ±0.013	0.913 ±0.011
	LWD	0.733 ±0.070	0.832 ±0.026	0.857 ±0.011	0.868 ±0.006	0.881 ±0.017	**0.920** ±0.010
	+ CoD	**0.738** ±0.093	**0.865** ±0.010	**0.870** ±0.017	**0.871** ±0.019	**0.885** ±0.007	0.913 ±0.013
Sent140	KD	0.580 ±0.039	0.597 ±0.042	0.645 ±0.023	0.690 ±0.035	0.752 ±0.011	**0.802** ±0.006
	+ CoD	**0.629** ±0.036	**0.640** ±0.048	**0.731** ±0.022	**0.754** ±0.017	**0.778** ±0.007	0.784 ±0.019
	LWD	0.581 ±0.041	0.593 ±0.039	0.665 ±0.027	0.708 ±0.029	**0.751** ±0.009	**0.785** ±0.019
	+ CoD	**0.628** ±0.034	**0.652** ±0.038	**0.706** ±0.016	**0.741** ±0.014	0.729 ±0.063	0.760 ±0.023

CoD achieves superior few-shot performance--outperforming standard distillation methods with as few as 8–128 samples--while **using only half the original data**, paired with their corresponding CFEs.

Conclusion

- Addresses few-shot distillation inefficiency by leveraging counterfactual explanations (CFEs).
- Statistical and Geometric motivations for CFE infusion.
- Empirically: CoD outperforms KD, LWD, and TED across benchmarks using only half the labeled data.
- CFEs turn explanations into data-efficient supervision, enabling faithful and robust few-shot distillation.



Poster: Wed 3rd December 1 pm -- 4 pm CST
Paper: https://arxiv.org/abs/2510.21631
Code: https://github.com/FaisalHamman/CoD

Thank you!