

Introduction

Background

- SciFact (Wadden et al., 2020): the standard benchmark dataset for scientific claim verification in NLI.
- 1,409 expert-annotated biomedical claims + evidence sentences drawn from a shared corpus of 5,183 abstracts.
- Pair labels: SUPPORT / NOT_ENOUGH_INFO / CONTRADICT.

Problem

- Claims split into train (809) or dev (300); but the abstract corpus is shared.
- Prior SciFact evaluations have not accounted for broad sentence-level exposure vs. exact pair-level memorisation distinction.

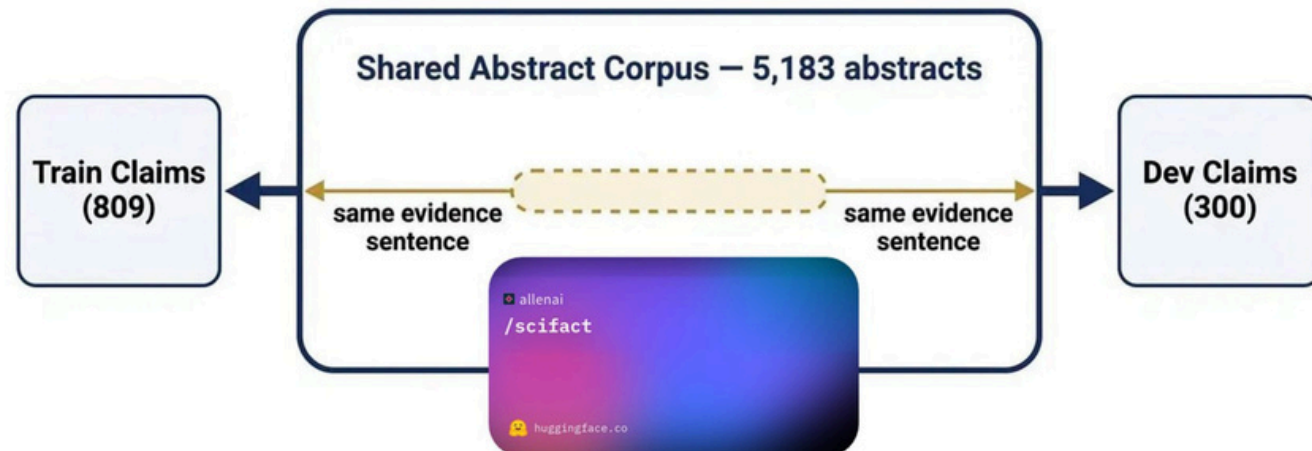


Fig. 1: Same evidence sentences can be cited by claims in both the train & dev splits, due to shared abstract corpus.

Aims & Objectives

- Measure both sentence-level (shared vocabulary) and exact pair-level (shared prediction targets) overlaps separately.
- Apply bilateral deduplication to remove any pair found in both splits.
- Report the true, leakage-corrected performance of a corrected split.

Discussion

- Since the leak is structural, any shared-corpus scientific NLI dataset including SciFact is susceptible.
- Bilateral deduplication fixes the leak without restructuring the underlying dataset.
- Published SciFact rankings, including RoBERTa-large, remain untested under this correction.

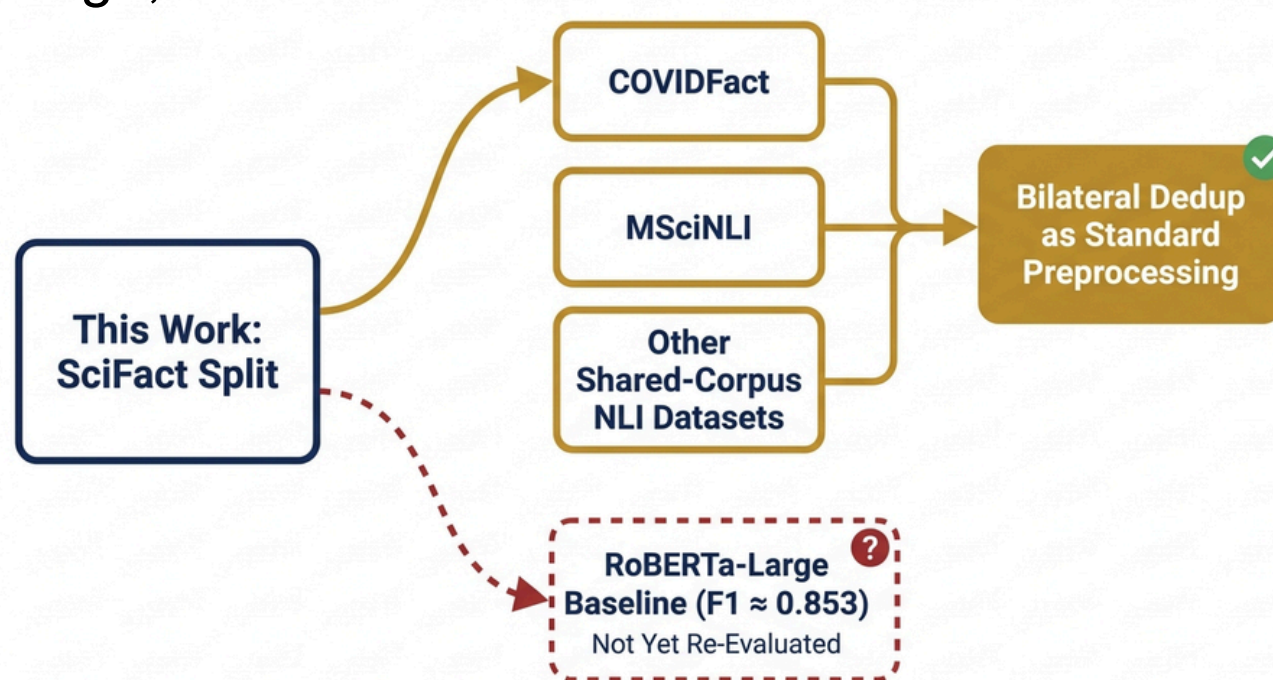


Fig. 5: Structural generalisation of leakage, and the untested SciFact leaderboard.

Presenters

Minoo La

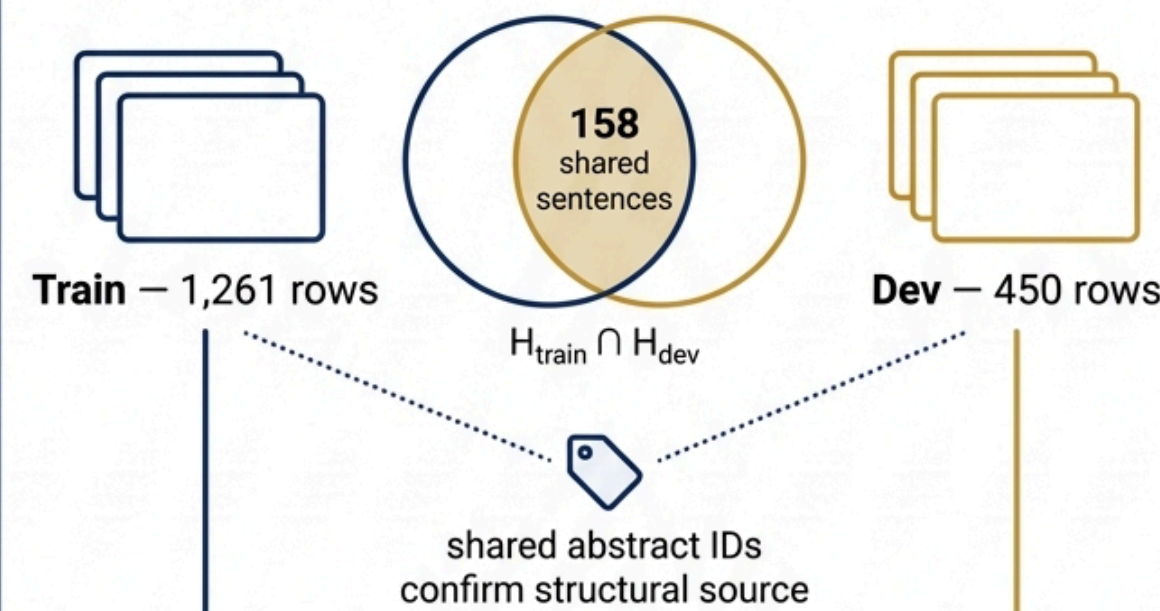
- Information Systems @ Korea University Business School
- lamino@korea.ac.kr | lamino.link

Sai Thet Hmuu

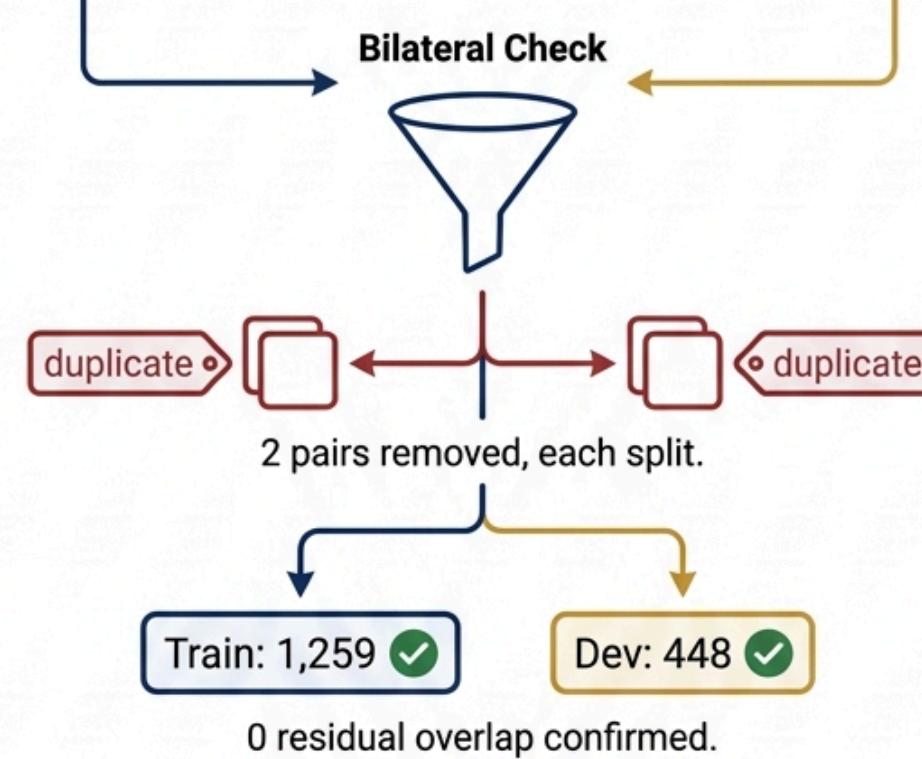
- School of Biosystems and Biomedical Sciences, Korea University
- pdkifoufrsl@korea.ac.kr

Methodology

1. Leakage Detection



2. Bilateral Deduplication



3. Model & Training

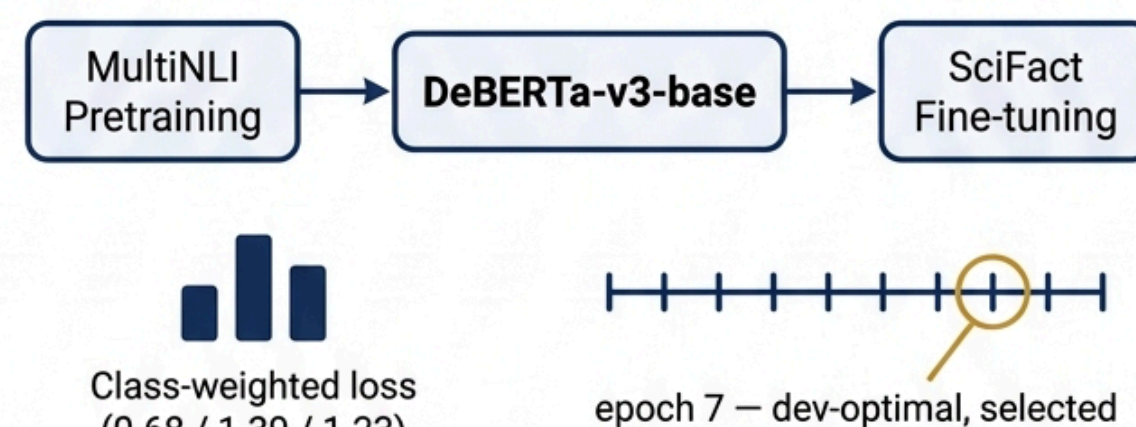


Fig. 2: Detection of sentence-level / pair-level overlaps between splits, bilaterally deduplicated to zero residual overlap, and fine-tuning the MultiNLI pre-trained DeBERTa-v3-base with the clean split.

Limitations

- Single training run and seed; the 1.6 point delta has not been bootstrapped or averaged.
- Correction measured only on the selected model (DeBERTa-v3-base, pretrained on MultiNLI, fine-tuned on the corrected SciFact split); other architectures were not considered.
- SciFact test labels are withheld, therefore test-set leakage cannot be assessed.

References

- Elangovan et al. (2021). *EACL 2021*, 1325-1335.
- He et al. (2023). *ICLR 2023*.
- Sadat & Caragea. (2024). *NAACL 2024*, 1610-1629.
- Wadden et al. (2020). *EMNLP 2020*, 7534-7550.
- Williams et al. (2018). *NAACL-HLT 2018*, 1112-1122.

Use of AI Tools

- ClaudeCode is used to edit certain parts of the codebase for the original project.
- FigureLabs is used to generate certain figures.



Results

Result 1

Sentence-level Overlap (Exposure) Vs. Bilateral Deduplication (True Contamination)

- 158 hypothesis sentences appear in both splits, touching 173 of 450 dev rows (38.4%). This number only indicates the model may have encountered similar text at some point.
- Pair-level checking, applied in both directions, removes exactly 2 (claim, evidence) pairs from each split, leaving 1,259 train and 448 dev rows with zero confirmed residual overlap. This number indicates a real memorisation target.

Result 2

Leakage Correction Lowers Reported Performance by 1.6 Points

- The model reaches macro F1 = 0.8641 and accuracy = 0.8705 on the corrected split, down from 0.8800 and 0.8844 on the standard split.
- The smaller number confirms a structural correction, and that the model had a genuine memorisation pathway available to it on these 2 rows.

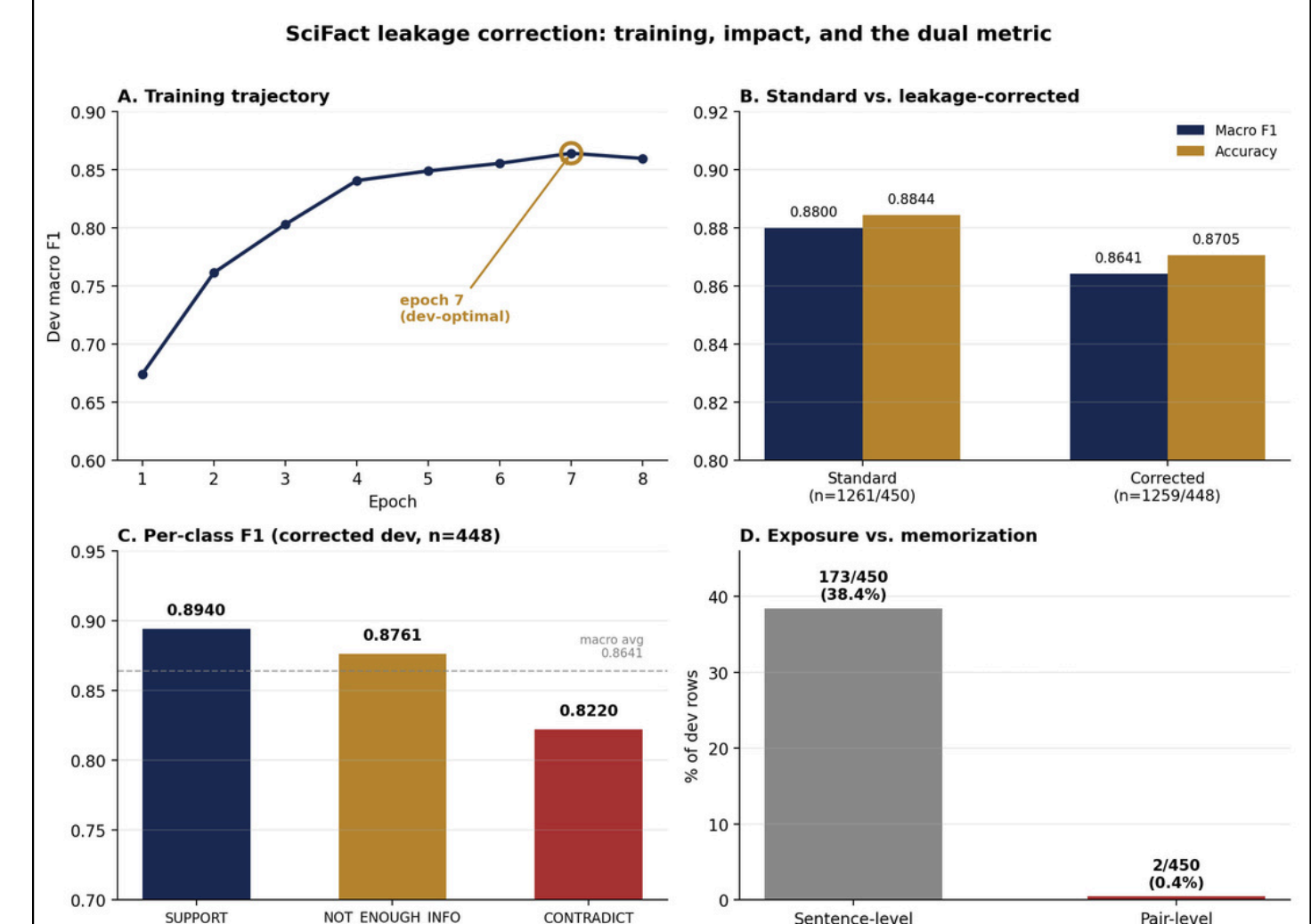


Fig. 3: Training curve, standard versus corrected performance, per-class F1, and the exposure versus memorisation contrast, all on the selected model.

Result 3

The Leak Is Concentrated in a Single Class

- All four bilaterally deduplicated rows, across both splits, belong to the NOT_ENOUGH_INFO class, consistent with the NEI construction procedure assigning the same placeholder sentence to claims citing the same abstract.

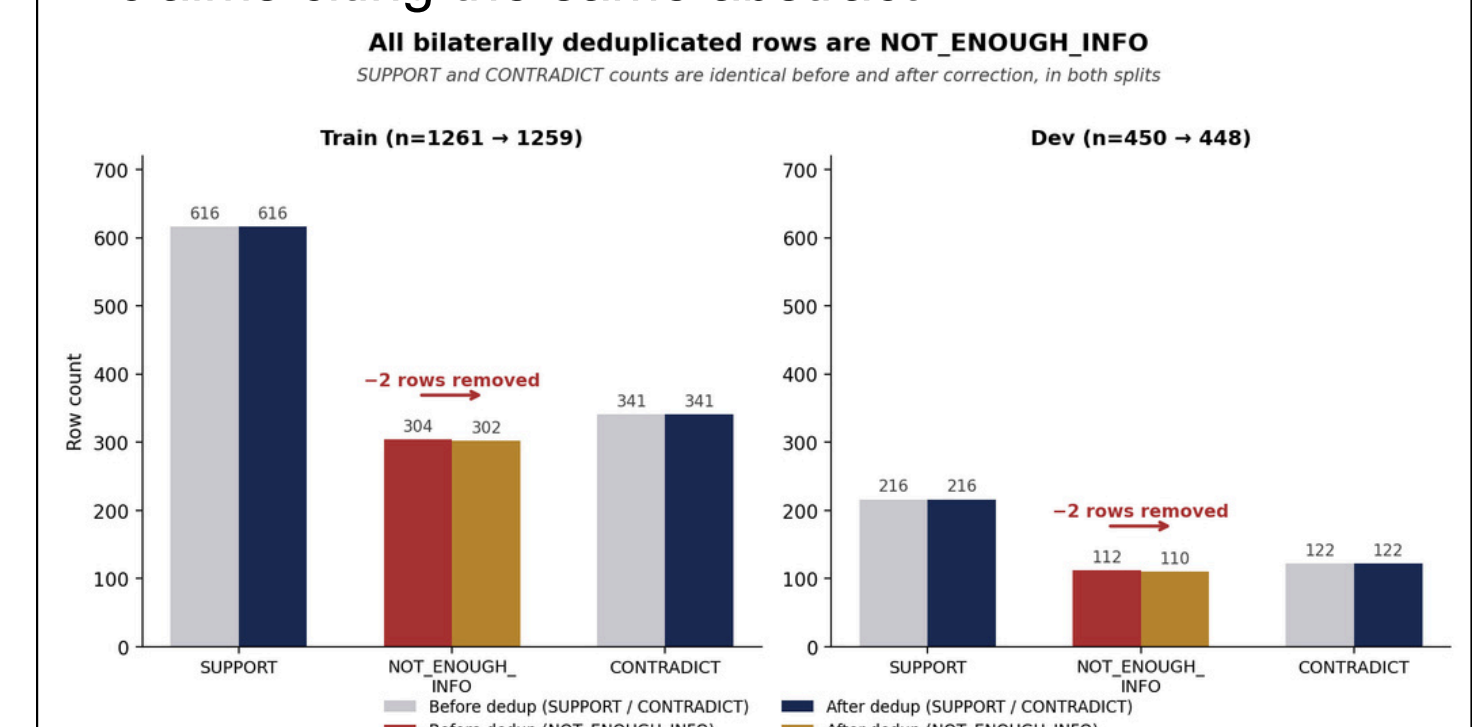


Fig. 4: Class-level breakdown of bilaterally deduplicated rows.

Acknowledgments

This work would not have been possible without the encouragement of Prof. Kyuhan Lee, Korea University Business School.