



Semi-Supervised Learning to Improve Detection of Factual and Causal Errors in Language Models' Explanations

Mingxuan Liu Han Yuan Yilin Wu Li Zhang Zheng Ma

Global Decision Science, American Express

{Mingxuan.Liu, Han.Yuan1, Yilin.Wu, Li.Zhang1, Zheng.Ma2}@amex.com

Background and Objective

- LM-generated explanations may contain factual errors or invalid causal reasoning.
- Reliable auditing requires fine-grained expert labels, which are scarce.
- Large volumes of unlabeled model-generated explanations are readily available.
- Objective 1:** Build a reasoning-point-level NLI verifier for the detection of factual and causal errors.
- Objective 2:** Use semi-supervised learning to improve verification with limited expert labels, and test its effectiveness across models and labeling budgets.

Methodology

- Reasoning-point decomposition:** Given a financial claim C and an LM-generated explanation $E = (r_1, \dots, r_m)$, LLM segments the explanation into individual reasoning points.
- Two binary NLI verification tasks:**
 - Factual verification: $C \Rightarrow r_i$; *Contradict* indicates a factual error.
 - Causal verification: $r_i \Rightarrow r_{concl}$; *Contradict* indicates a causal error.
- NLI verifier training:** We compare supervised fine-tuning, pseudo-labeling, and FreeMatch. FreeMatch uses adaptive confidence thresholds and consistency regularization to better leverage unlabeled explanations.
- Labeling-budget evaluation:** We vary the labeled-data budget 100%, 75%, 50%, 25%, 10% while keeping the unlabeled dataset fixed.
- Verifier architectures: Experiments are conducted with both BART (encoder-only) and Gemma-2-2B (decoder-only).

Experiments

Factual Error Detection

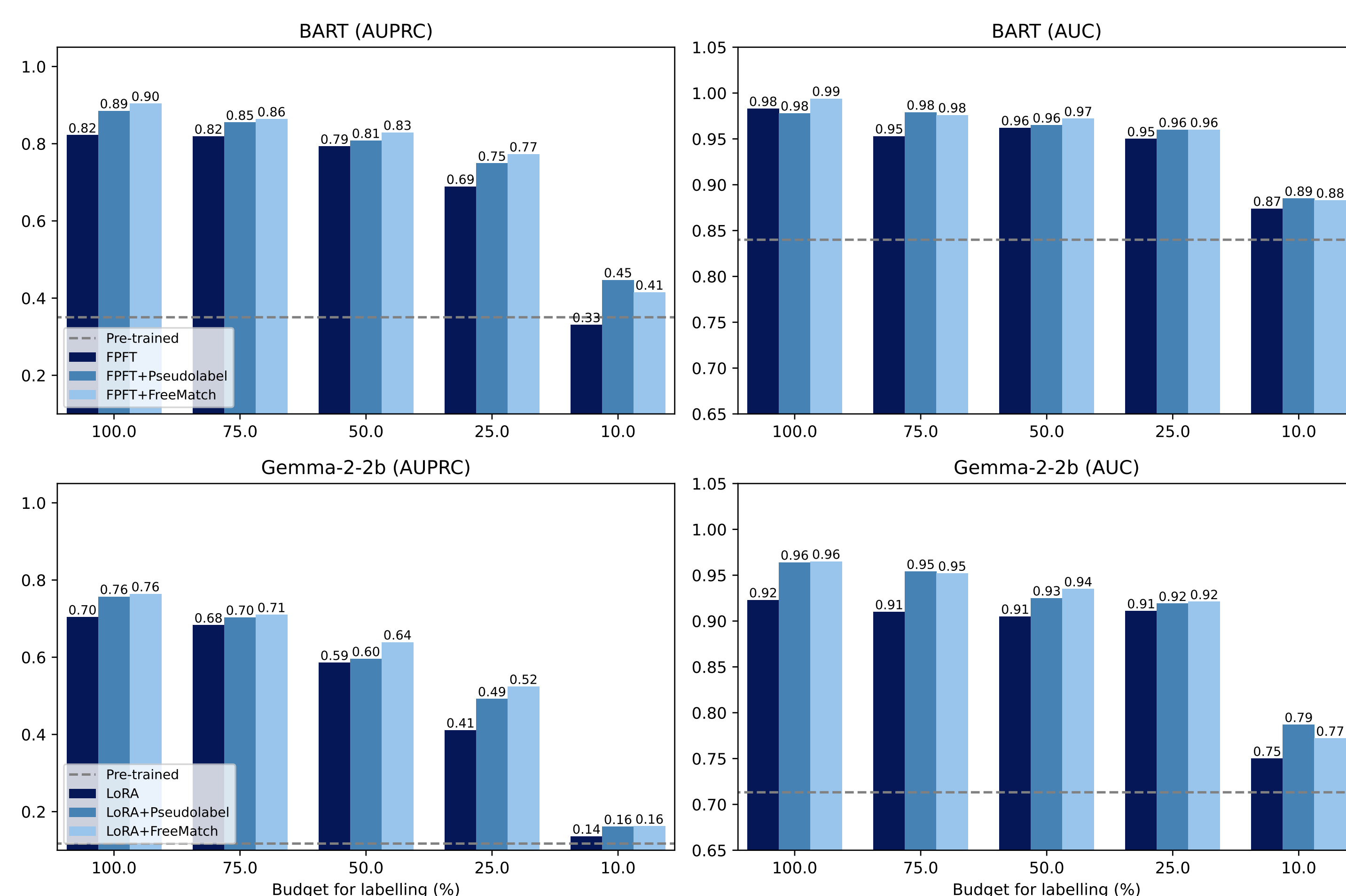


Figure 1: Factual error detection across labeling budgets for BART and Gemma-2-2B.

- SSL improves over both the pre-trained baseline and supervised fine-tuning for both verifiers.
- Gains show diminishing returns as labeling budget increases.
- FreeMatch is generally as good as or stronger than pseudo-labeling, especially at moderate-to-high budgets.

Causal Error Detection

- SSL also improves causal-error detection over both pre-trained and supervised baselines.
- At a 50% labeling budget, Gemma-2-2B causal AUPRC increases from 0.41 (LoRA) to 0.81 (Pseudo-labeling) and 0.8 (FreeMatch).

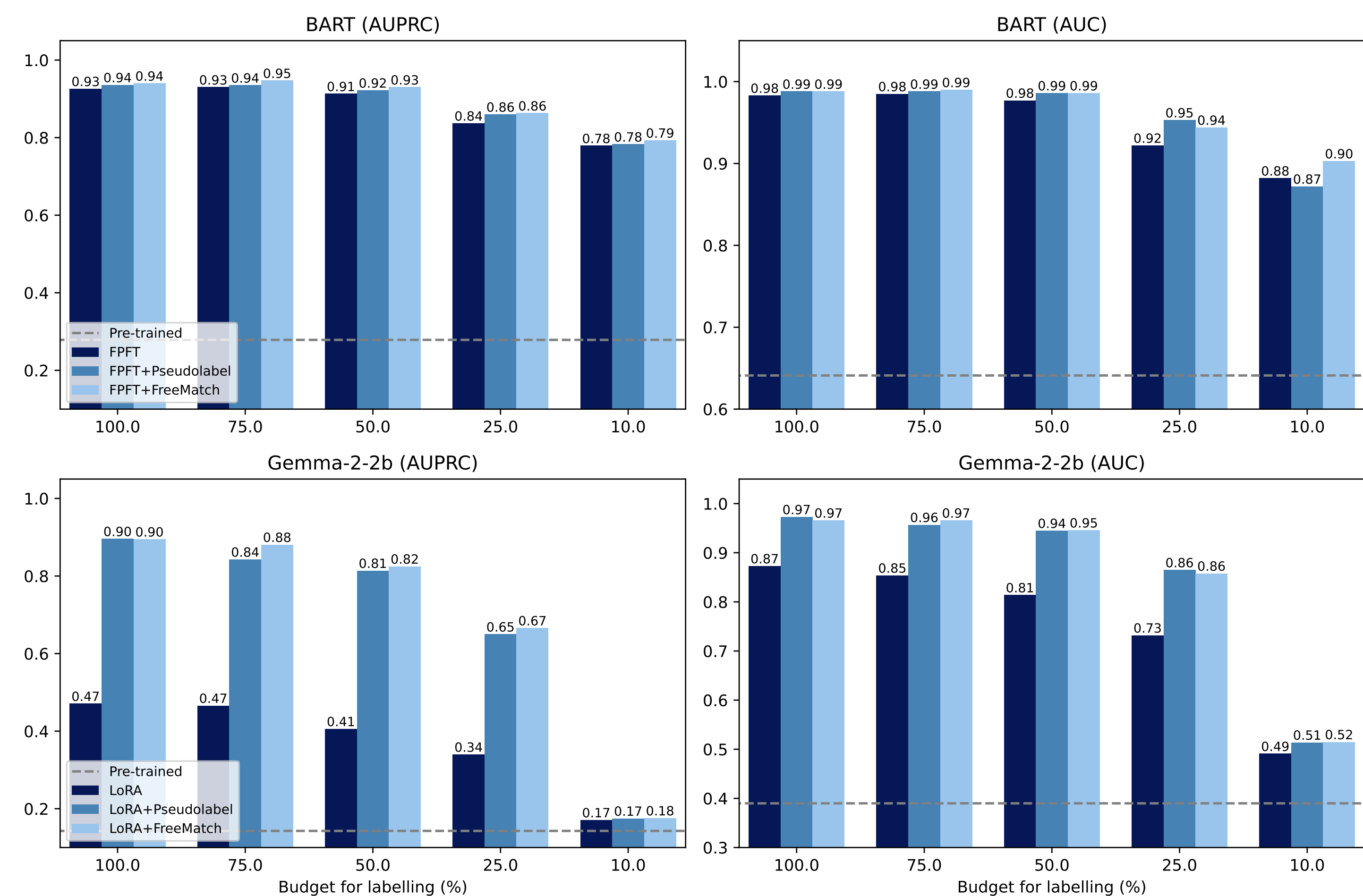


Figure 2: Causal error detection across labeling budgets for BART and Gemma-2-2B.

Unlabeled-Data Utilization

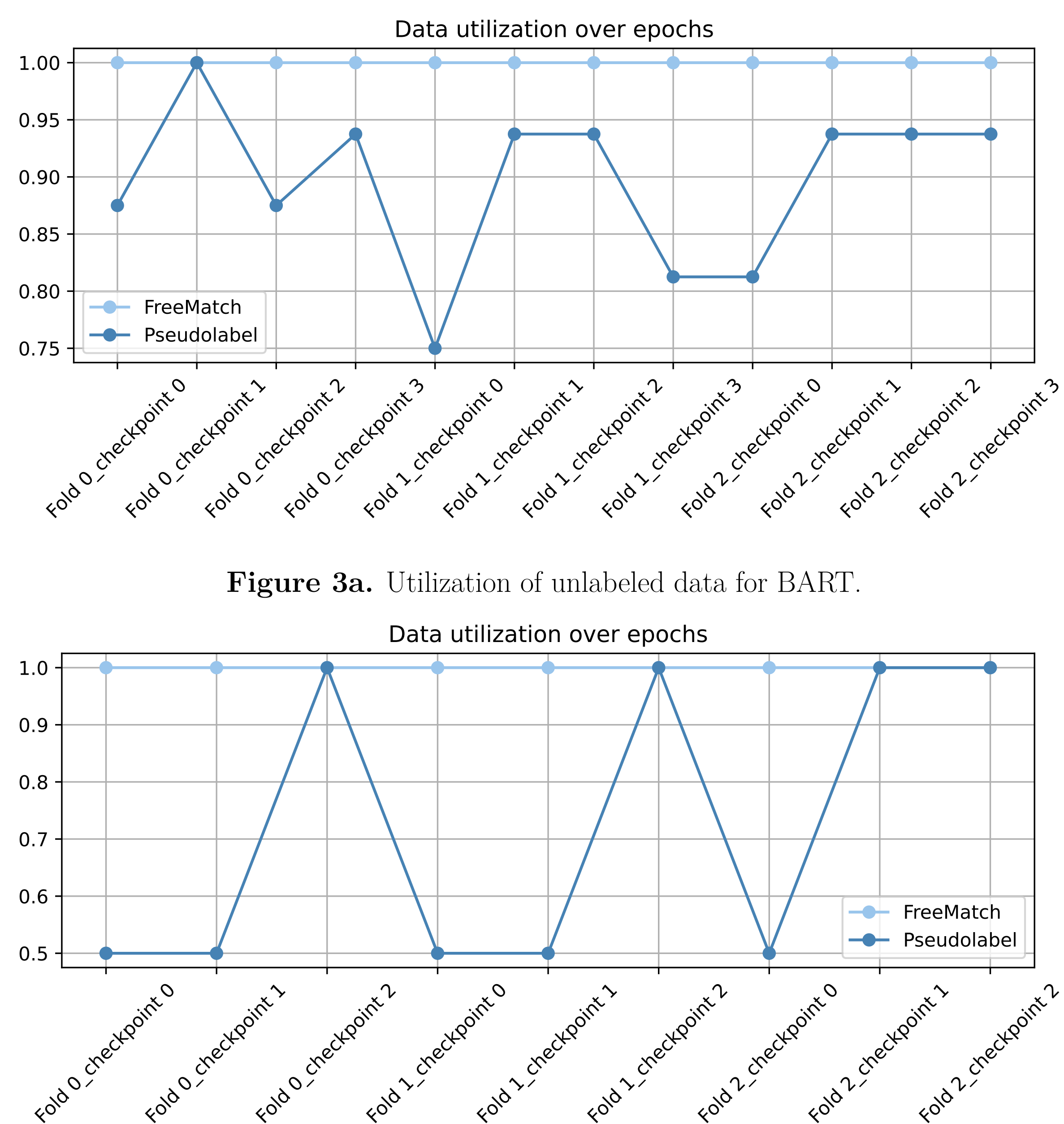


Figure 3a. Utilization of unlabeled data for BART.

Figure 3b. Utilization of unlabeled data for Gemma-2-2B.

- FreeMatch stays close to full utilization across training for both verifiers.
- Pseudo-labeling fluctuates more strongly, especially for Gemma-2-2B, indicating less stable use of unlabeled data.

Disclaimer

This paper is provided solely for informational purposes as an academic contribution by the authors to the research community and does not represent, reflect, or constitute the views, policies, positions, or practices of American Express or its affiliates. Nothing in this paper should be cited or relied upon as evidence of, or support for, the business views, policies, positions, or practices of American Express or its affiliates.

References

- Lee, D.-H. 2013. *Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*. ICML Workshop on Challenges in Representation Learning.
- Wang, Y. et al. 2023. *FreeMatch: Self-Adaptive Thresholding for Semi-Supervised Learning*. ICLR 2023.
- Yuan, H., Zhang, L., and Ma, Z. 2025. *Exploring the Reliability of Self-Explanation and Its Relationship with Classification in Language Model-Driven Financial Analysis*. arXiv:2503.15985.

Takeaways

- SSL consistently improves factual and causal error detection in LM-generated explanations.
- FreeMatch is generally more robust than simple pseudo-labeling, with better control of pseudo-label noise and unlabeled-data utilization.
- More labeled data improves NLI verification, but with diminishing returns.