ICML
International Conference
On Machine Learning

arXiv

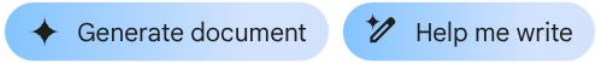
GitHub

ACTG-ARL: Differentially Private Conditional Text Generation with RL-Boosted Control

Yuzheng Hu, Ryan McKenna, Da Yu, Shanshan Wu,
Han Zhao, Zheng Xu, Peter Kairouz

ICML 2026

Differentially Private Synthetic Text



Text data are important in fueling various applications

User data are private

DP synthetic text

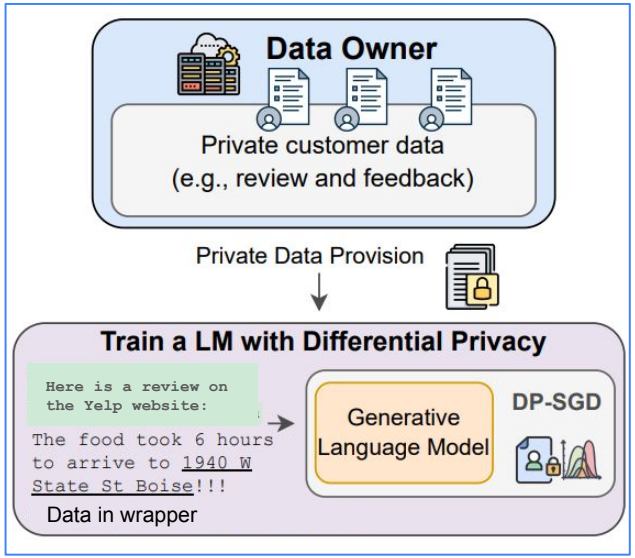
One-time privacy cost

Any downstream task

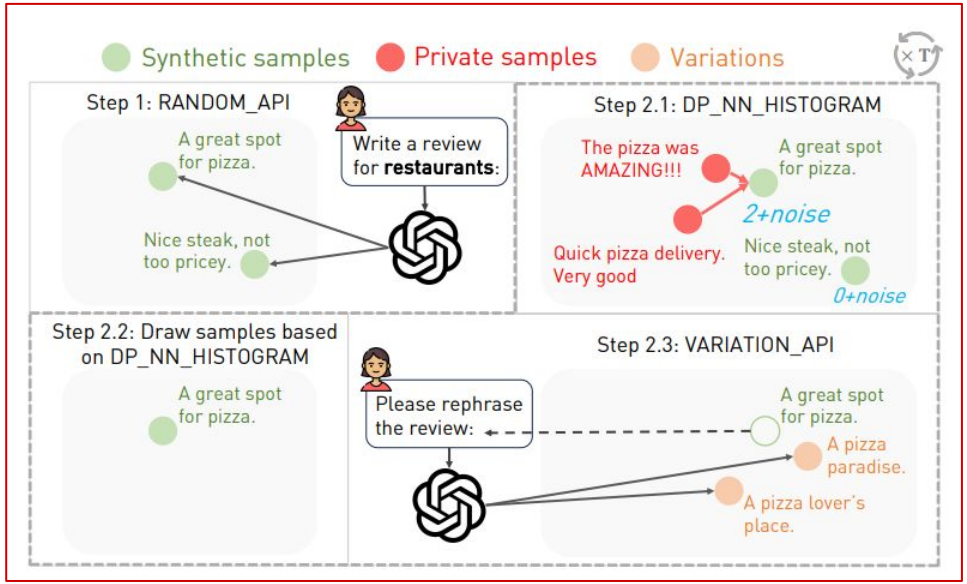
No need to modify the original data processing pipeline

Prior Work

Private fine-tuning (DP-FT)
(Yue et al., 2023)



Private evolution (PE)
(Xie et al., 2024)



Yue, Xiang, et al. "Synthetic text generation with differential privacy: A simple and practical recipe." ACL 2023.
Xie, Chulin, et al. "Differentially private synthetic data via foundation model apis 2: Text." ICML 2024.

Drawbacks of Current Approaches

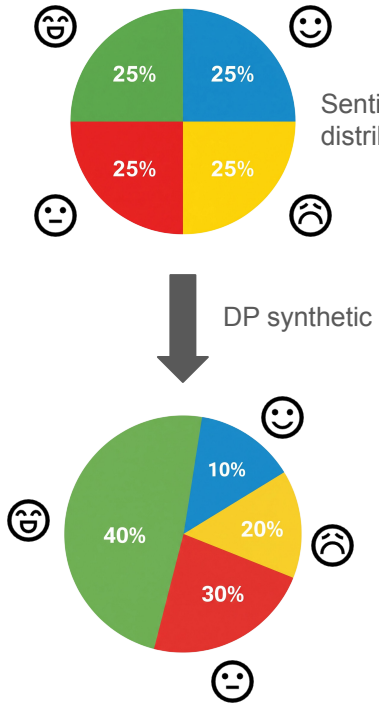
Limited Controllability

Write an email that represents the following schema:

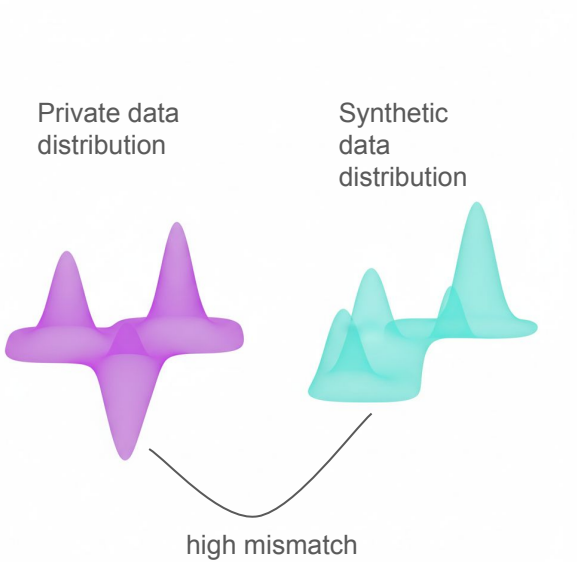
```
{'token_counts': 81,
 'sentiment': 'Positive',
 'tone': 'Direct',
 'purpose': 'Approval Request',
 'urgency': 'Normal',
 'formality': 'Neutral',
 'cta': 'Explicit',
 'attachments': False,
 'email_type': 'Conversational',
 'sender_relation': 'Internal Colleague',
 'topics': ['Hiring', 'Human Resources', 'Executive Approval']}
```

DP Generator

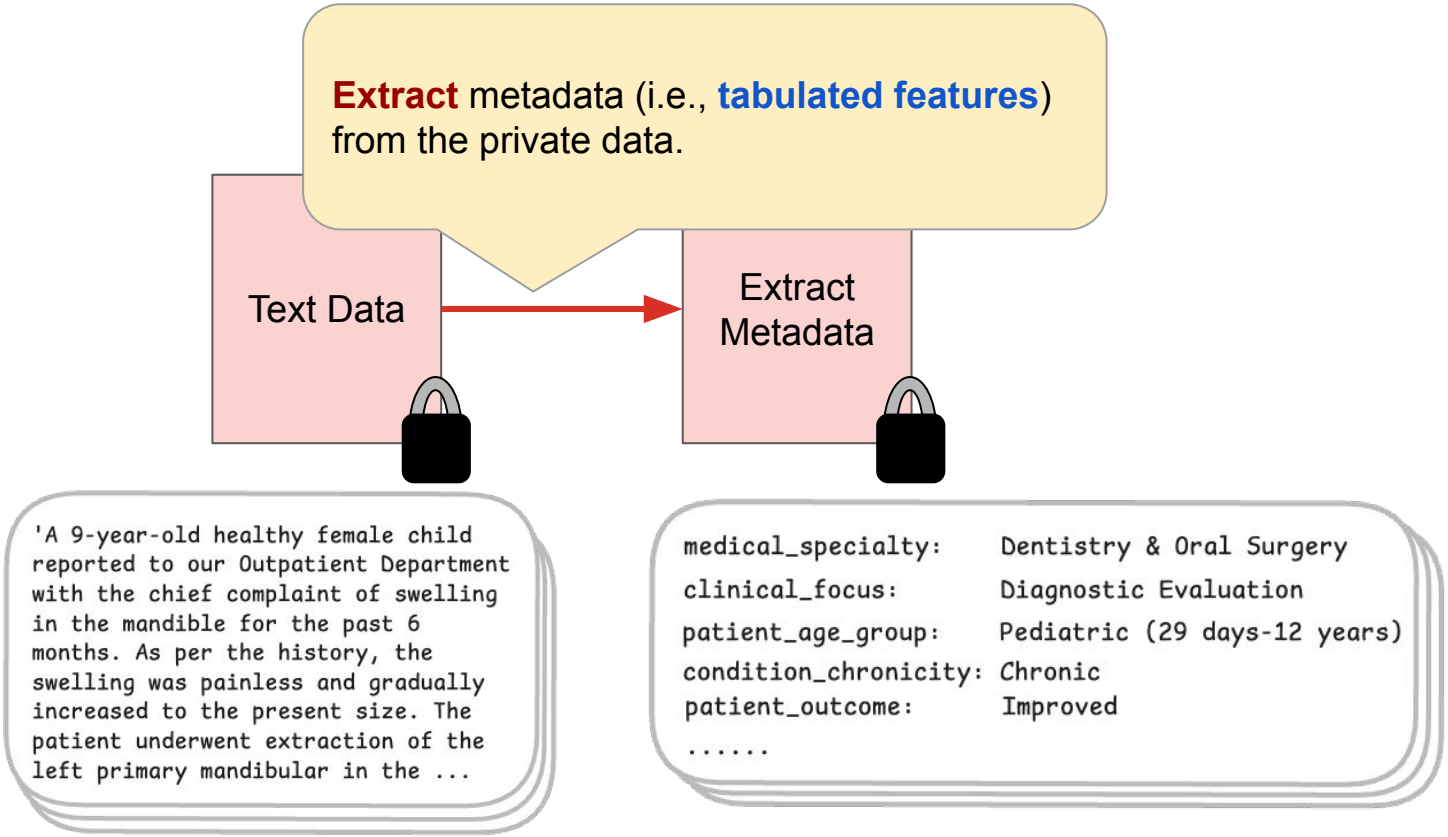
Imbalanced Feature Representation



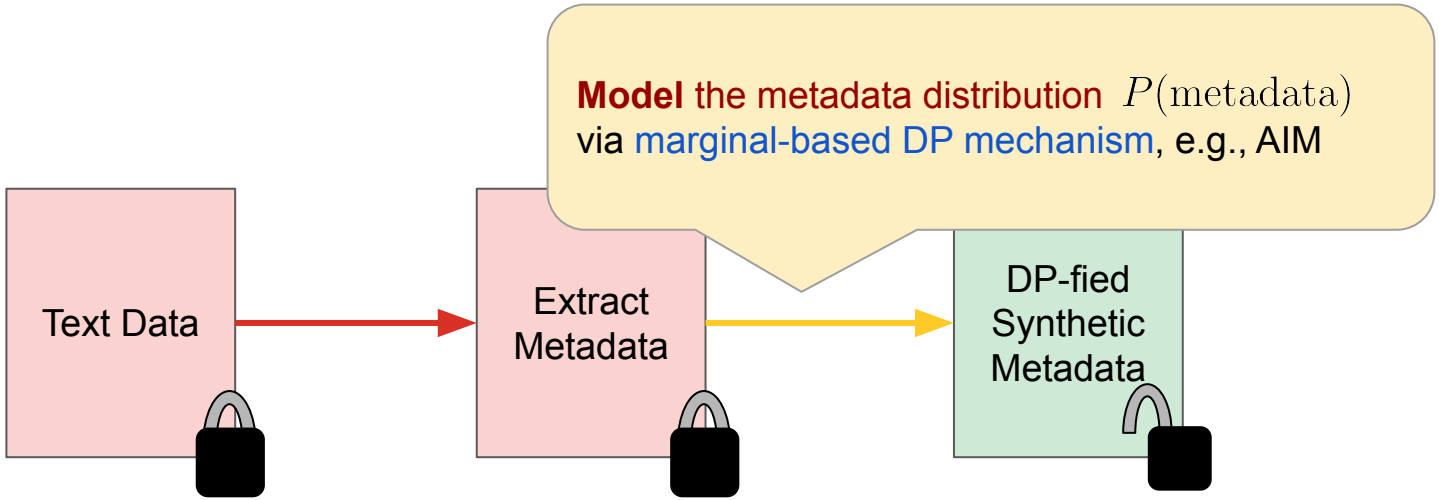
Low Data Quality



Schema Design & Metadata Extraction

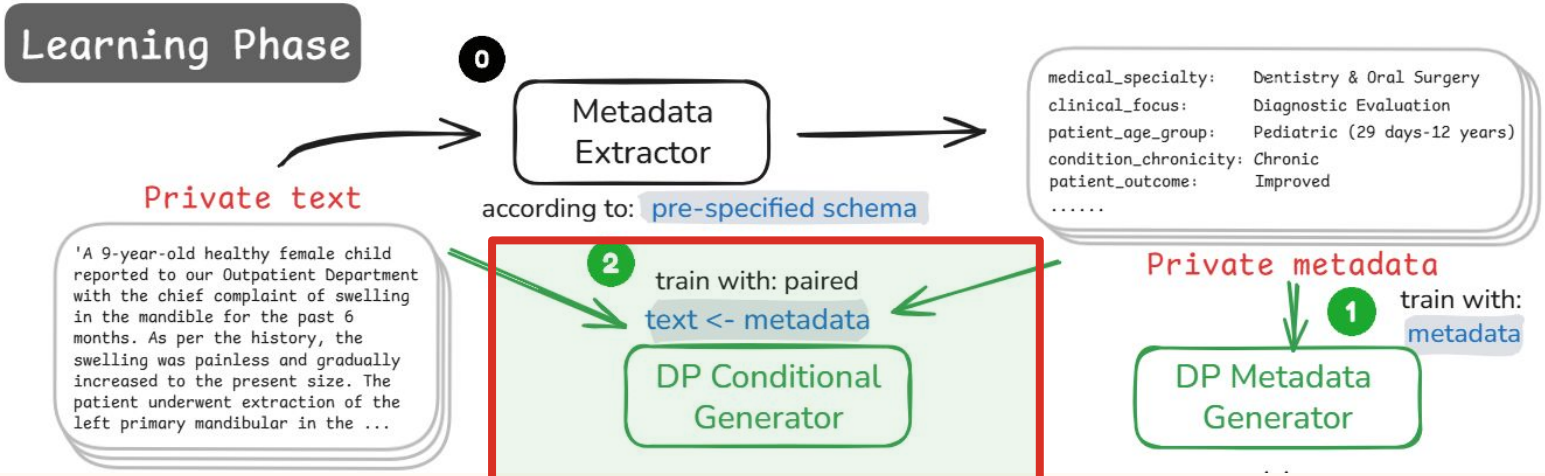


Privately Fitting the Metadata Distribution



McKenna, Ryan, et al. "AIM: An adaptive and iterative mechanism for differentially private synthetic data." *PVLDB* 15.11 (2022): 2599.

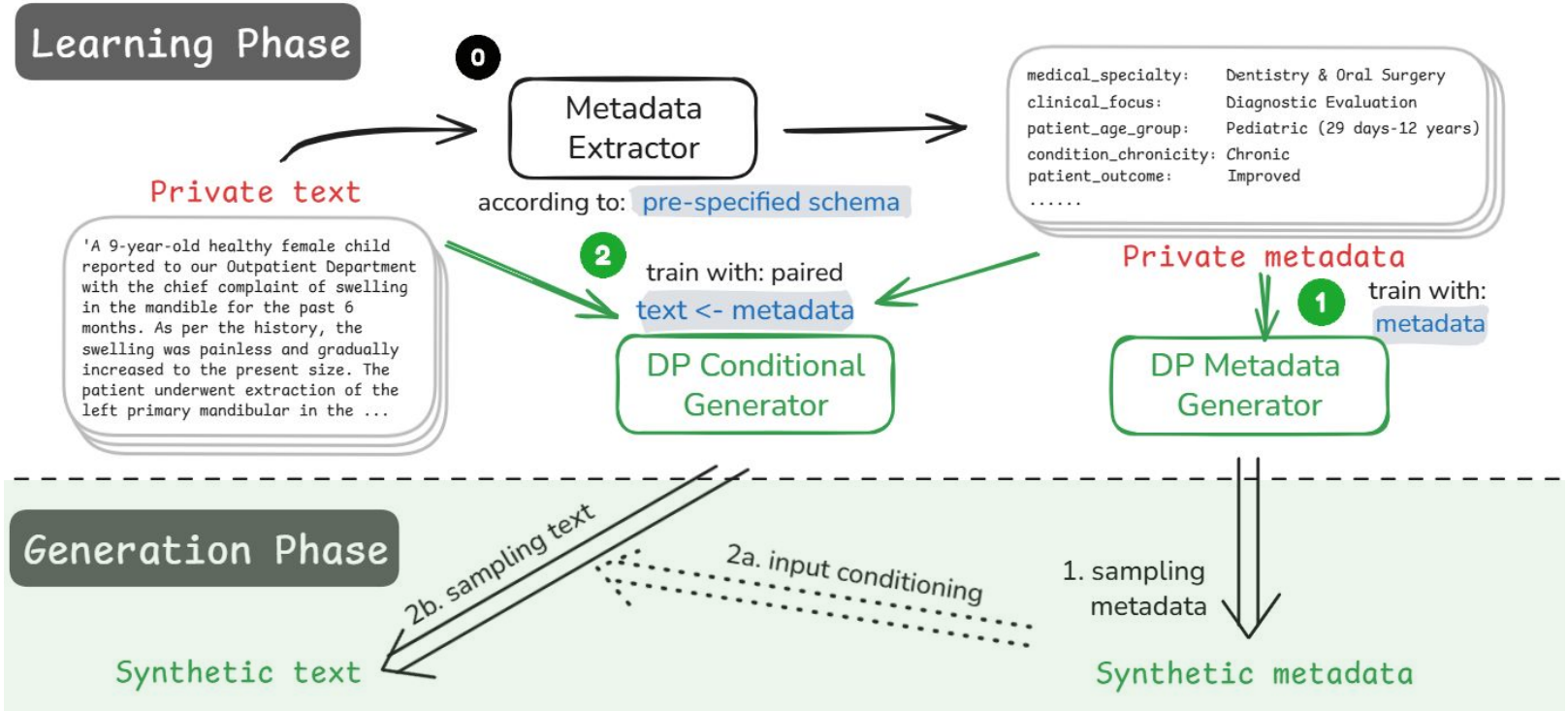
Conditional Fine-Tuning



Fine-tuning template

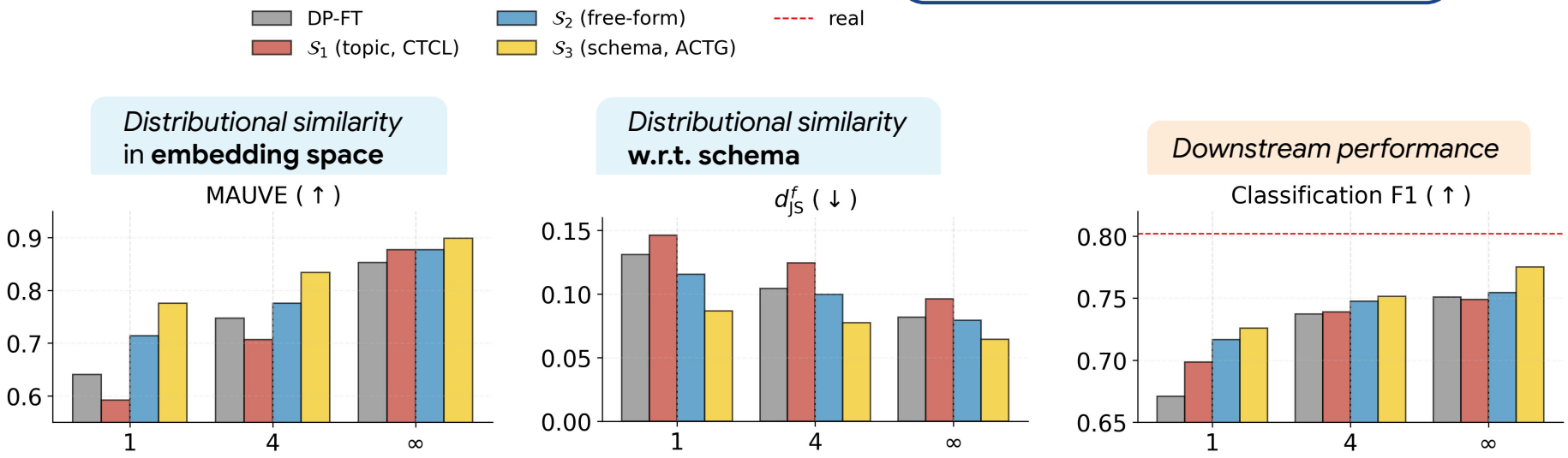
```
<start_of_turn>user
Please generate a detailed clinical note based solely on the below JSON summary, covering the patient's visit, medical history, symptoms, administered treatments, and outcome of the intervention.'
{{feature}}
<end_of_turn>
<start_of_turn>model
{{text}}
```

Conditional Generation



Results

Model: Gemma-3-1B-PT
Dataset: bioRxiv
Privacy budget: 1, 4, ∞



ACTG: **A**ttribute-**C**onditioned **T**ext **G**eneration

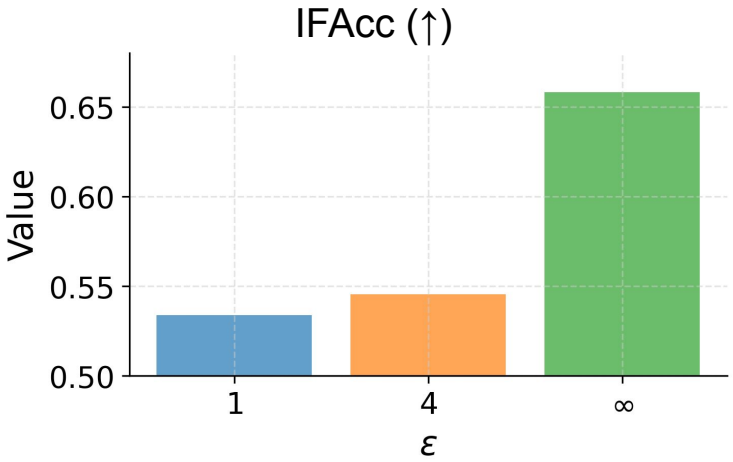
Metadata + AIM feature generator + DP-FT conditional generator

A new **state of the art** in DP synthetic text generation

Beyond DP Synthetic Text: *Controlled* Generation

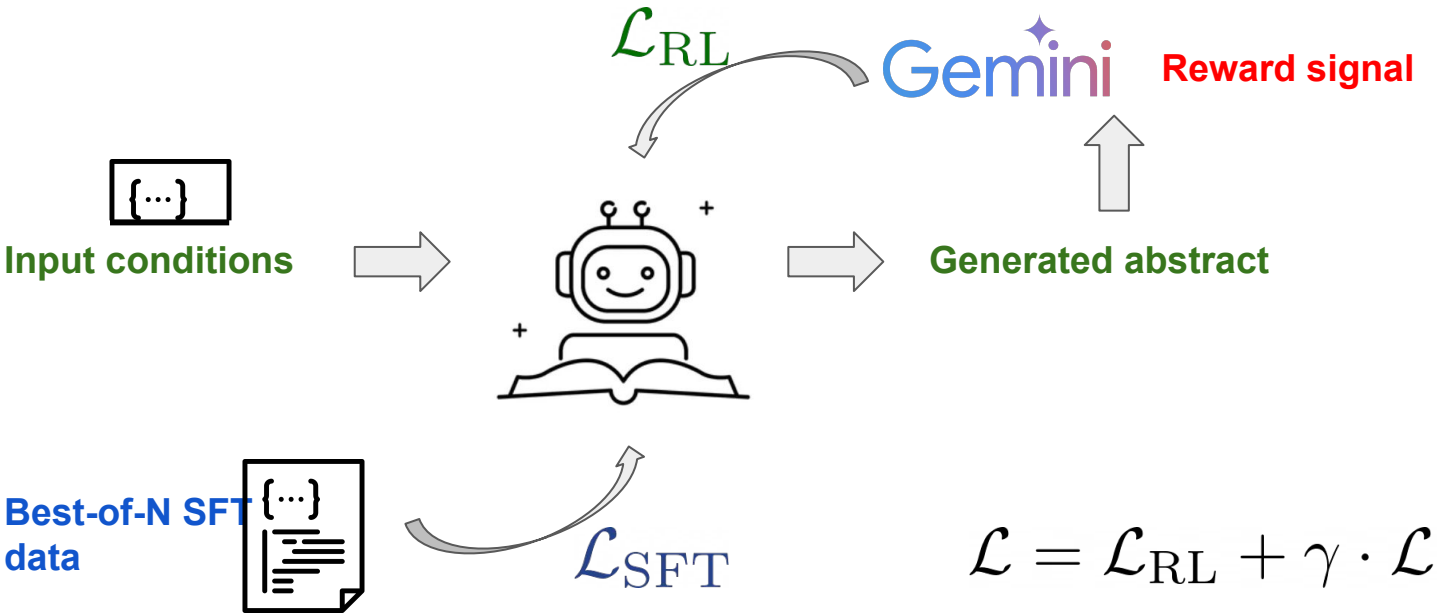
Instruction following accuracy (IFAcc):

The percentage of correct fields per sample, averaged over all samples

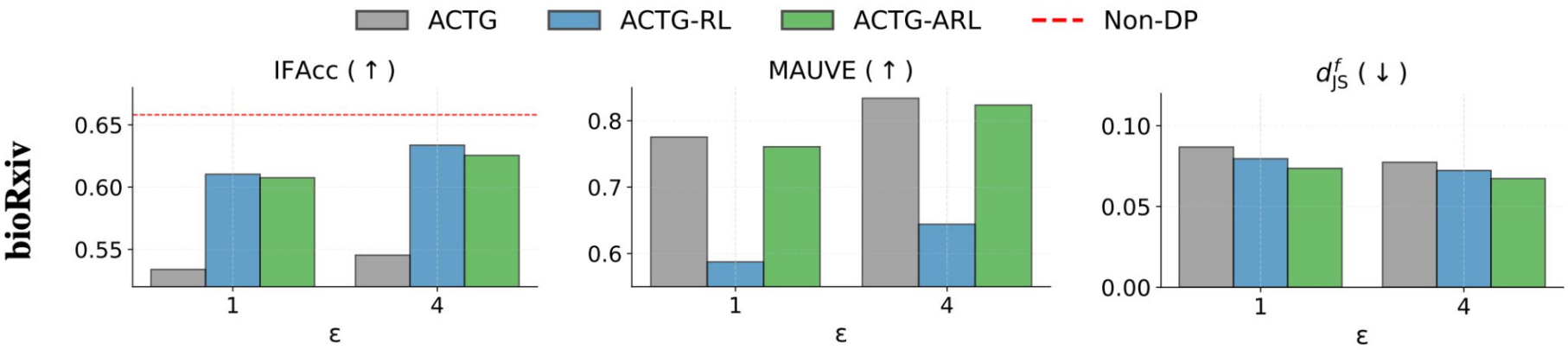


Incorporating DP in SFT significantly **hurts** the model's instruction following capabilities

A Post-Training Recipe: Anchored RL



Results



A conditional text generator with **improved control**

Thanks!

arXiv

GitHub