

Aligning LLMs with Biomedical Knowledge using Balanced Fine-Tuning

Zhenchao Tang^{*1}, Fang Wang^{*†1}, Haohuai He¹, Jiale Zhou¹, Tianxu Lv¹, Jun Zhu¹, Shouzhi Chen¹, Minghao Yang¹, Yu Wang¹, Jiayang Wu¹, Yidong Song², Yaokun Li³, Jiehui Huang⁴, Jun Zhou⁵, Bing He¹, Jianhua Yao^{†1}

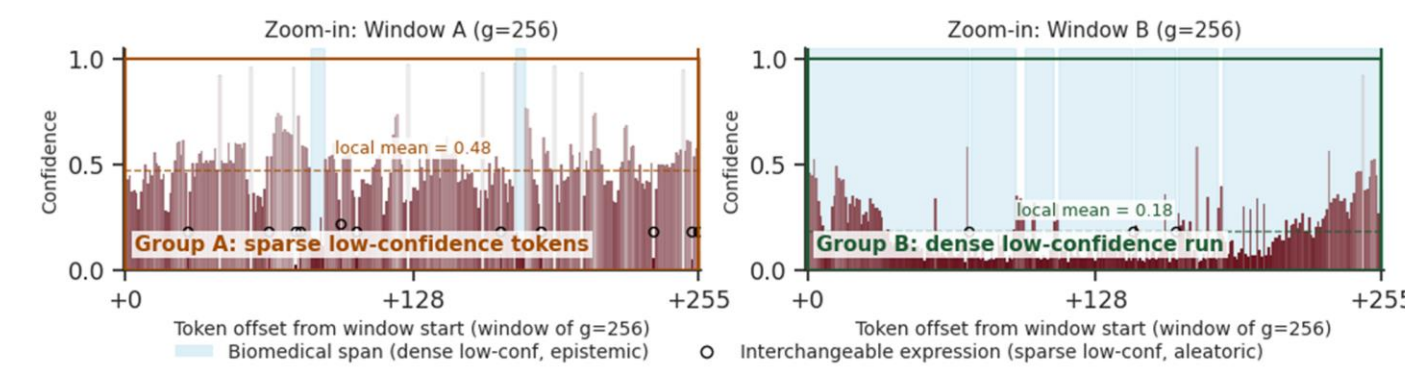
¹Tencent AI for Life Sciences Lab ²Nankai University ³Sun Yat-sen University ⁴HKUST ⁵Tsinghua University

^{*}Equal contribution [†]Correspondence: avonwanghit@gmail.com, jianhua.yao@gmail.com

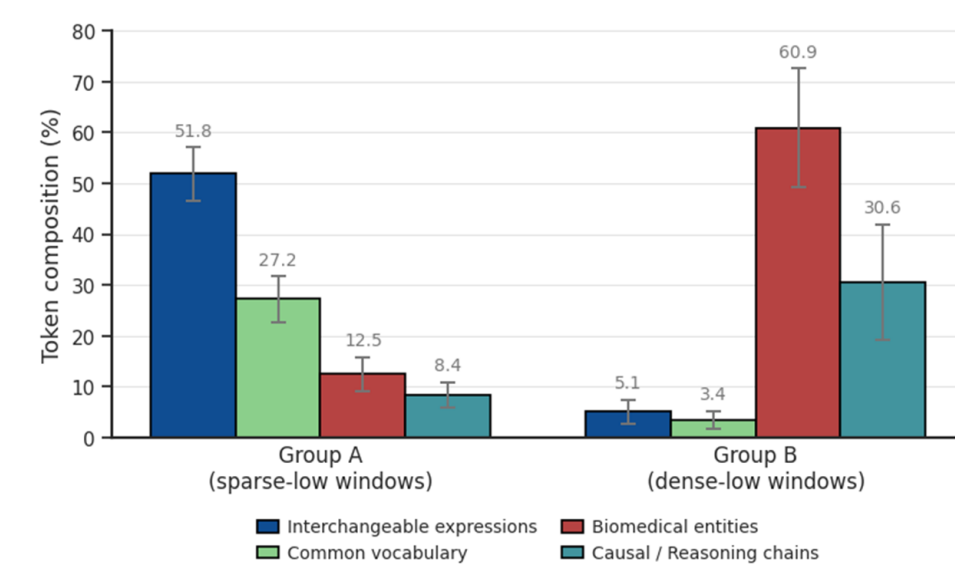


1 Challenge: useful knowledge is difficult

Confidence-based training can suppress unfamiliar biomedical words. An isolated difficult word and a whole difficult passage need different treatment.



2 Motivation: learn difficult passages



Grouped difficult passages contain more biomedical entities (60.9% vs 12.5%) and reasoning content (30.6% vs 8.4%).

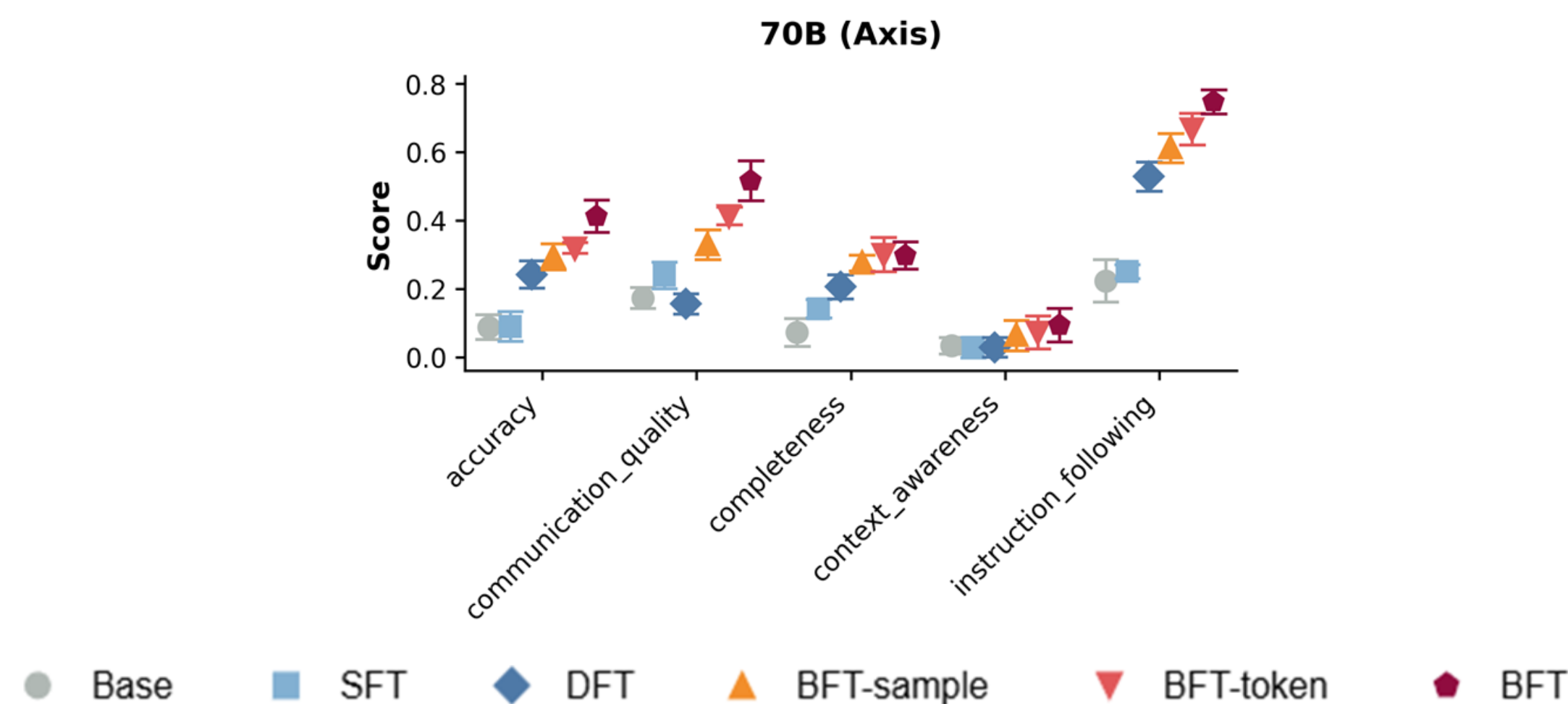
Preserve useful difficulty while retaining broad knowledge.

These associations motivate BFT. They do not identify every uncertain word as useful knowledge.

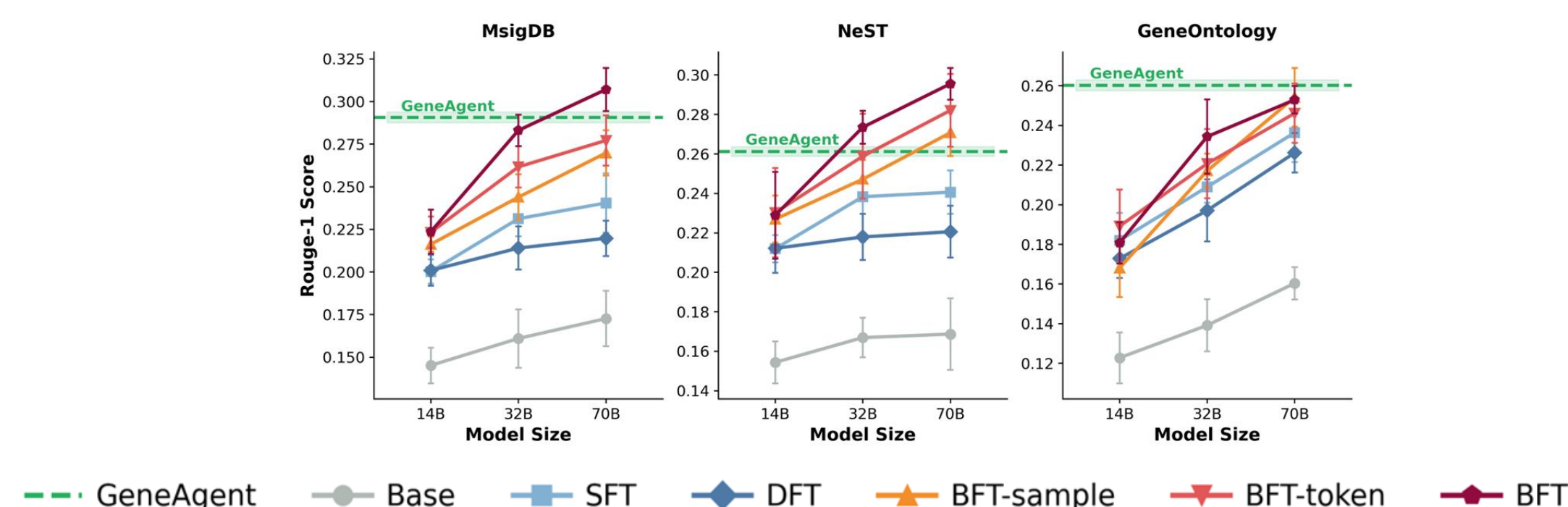
4 Experimental results: biomedical gains with less general forgetting

Stronger biomedical reasoning

HealthBench Hard (70B): medical competence across five evaluation dimensions.

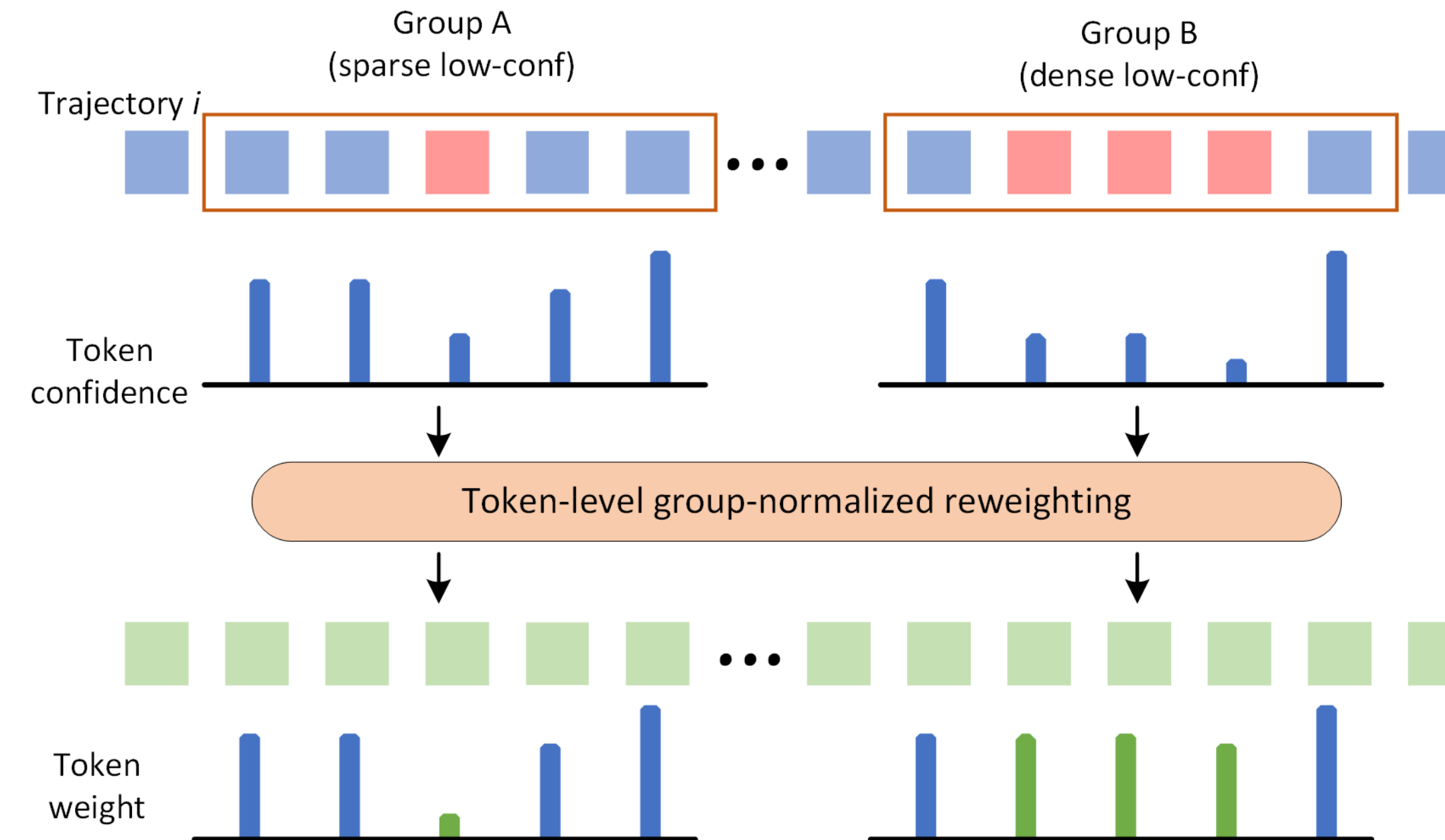


Biological process reasoning



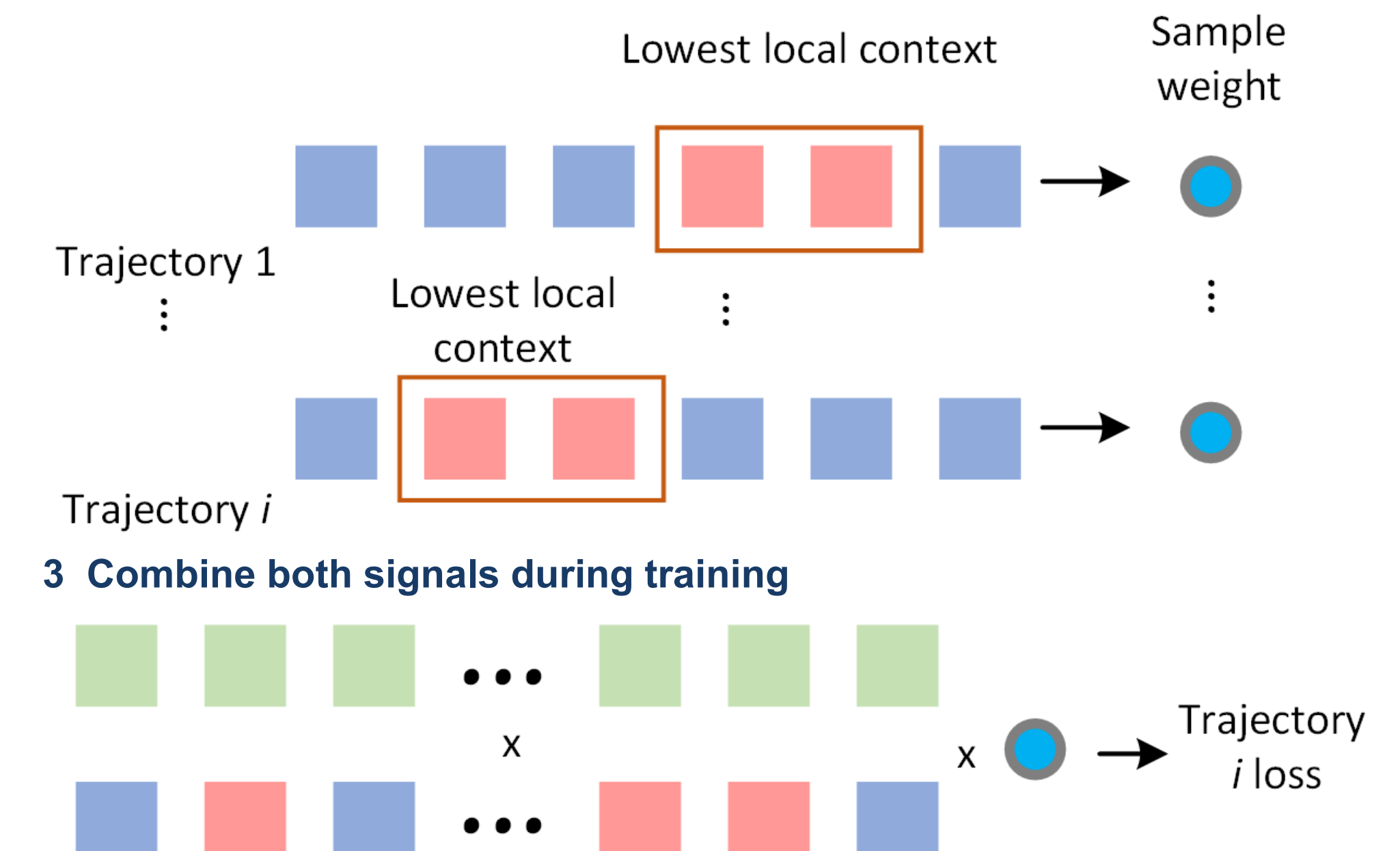
3 Method: Balanced Fine-Tuning

1 Keep useful words in difficult passages



Use the model's own confidence; cap example weights. No reward model or external uncertainty estimator.

2 Prioritize examples with hard passages

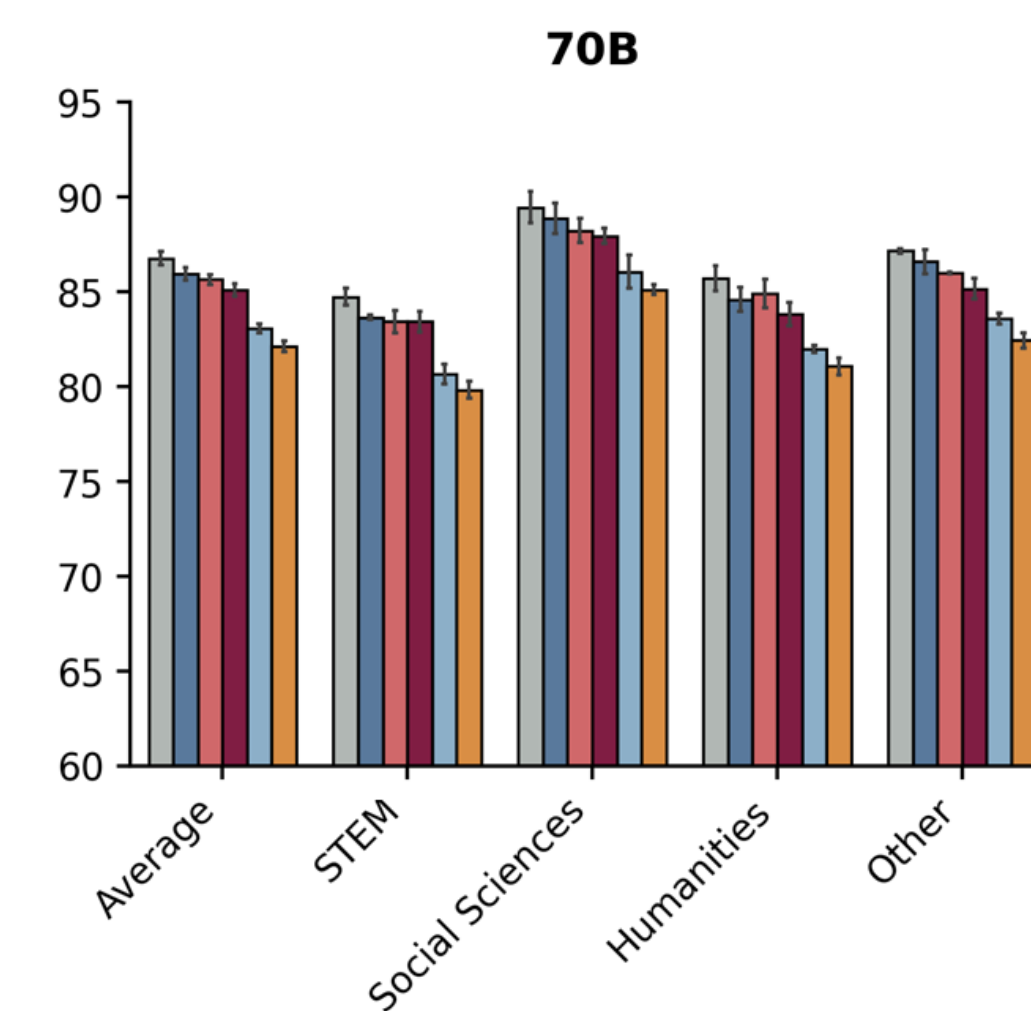


3 Combine both signals during training

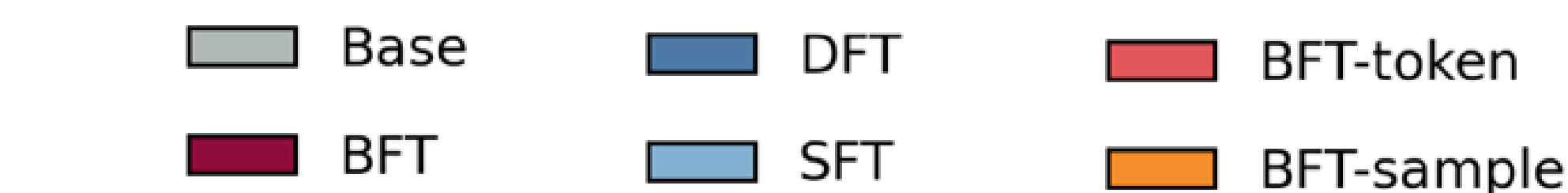
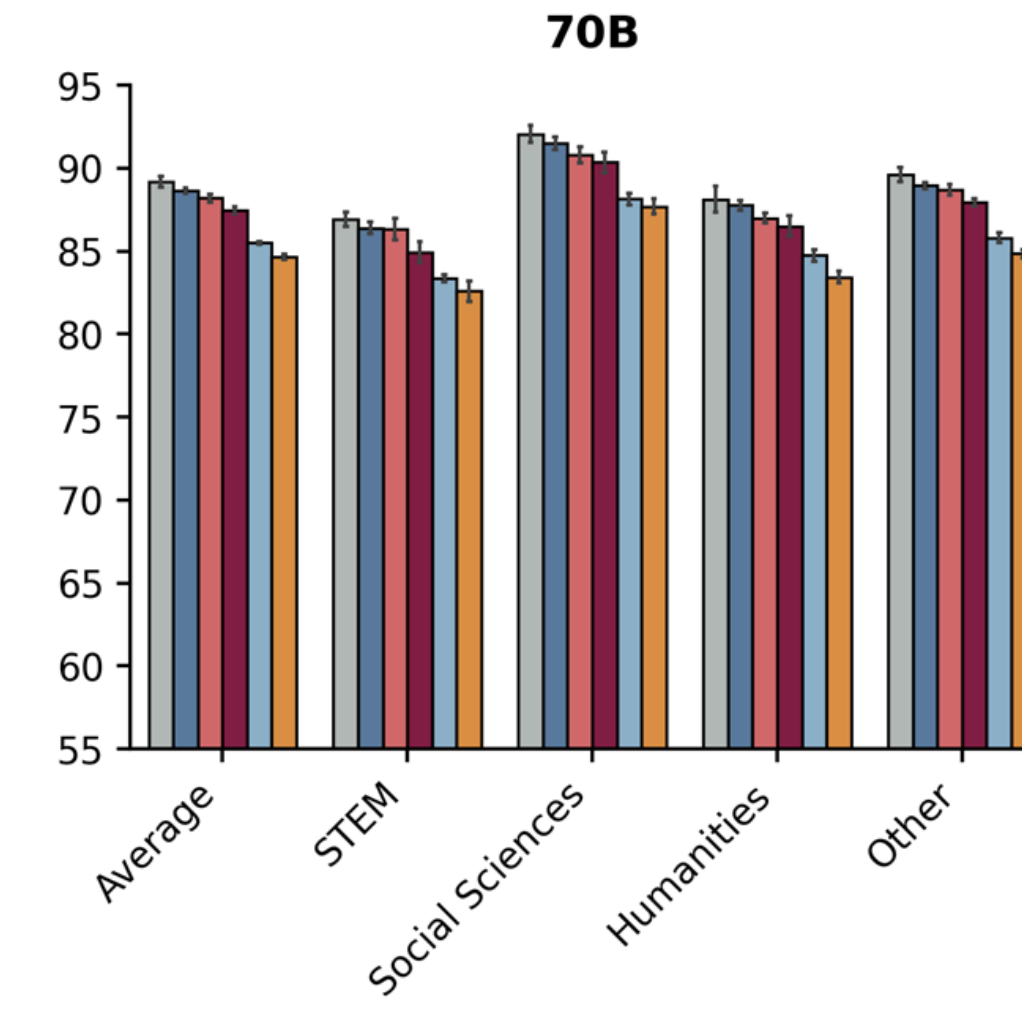
Preserving general capability

MMLU and CMMLU (70B): BFT preserves more general knowledge than SFT while improving biomedical skills.

MMLU



CMMLU

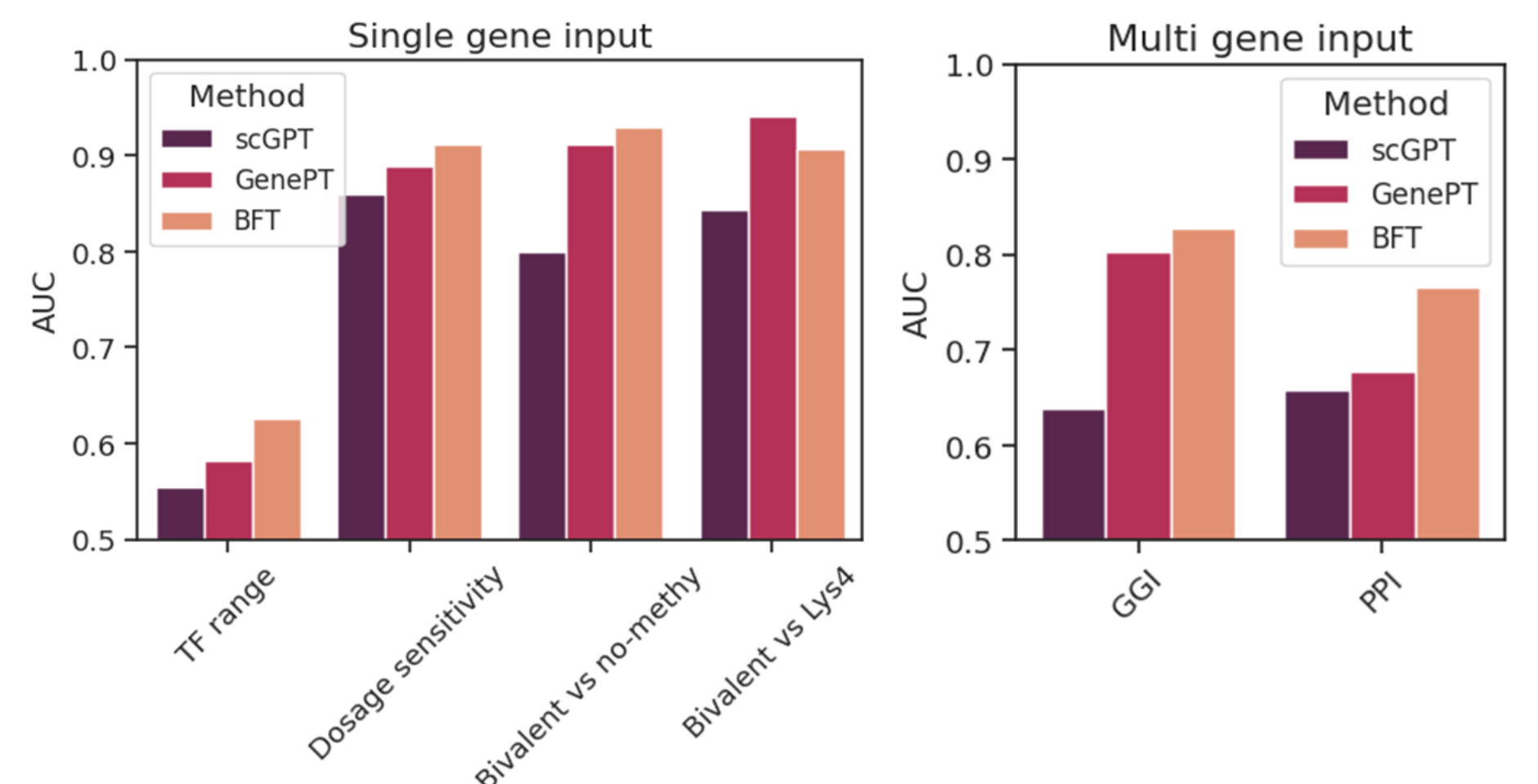


Both benchmarks show less forgetting than SFT and sample-only BFT. Some forgetting remains. DFT forgets least, while BFT offers stronger biomedical adaptation.

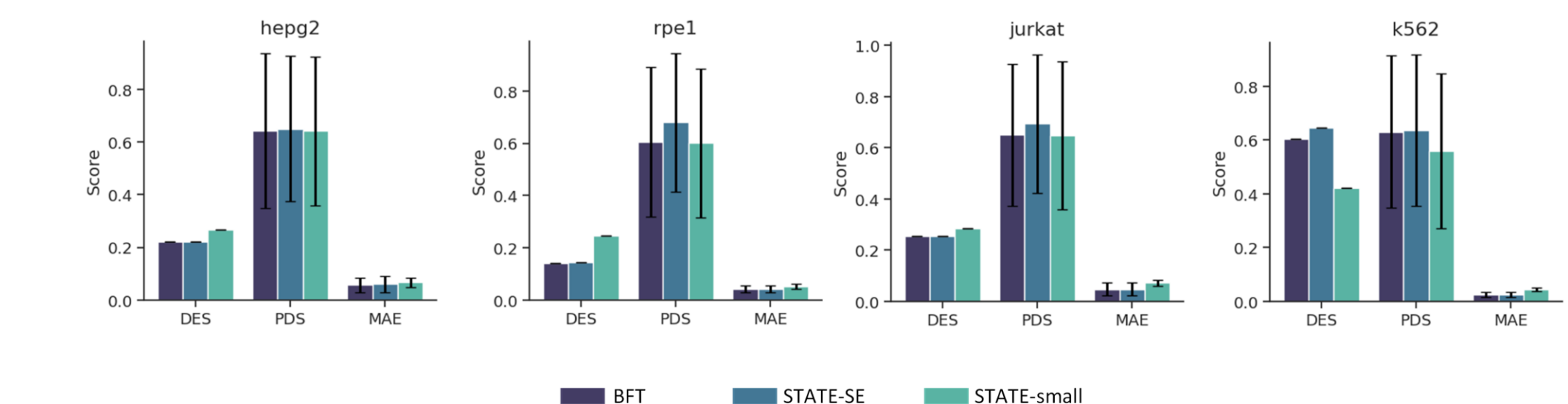
STEM adaptation demonstrated in biomedicine.

Useful scientific representations

Gene descriptions become inputs for gene prediction and single-cell response modeling.



Mean gene score: BFT 0.829 vs GenePT 0.802. GenePT remains higher on Lys4. Both use the same text encoder.



Comparable STATE results across four datasets. Profile generation is zero-shot here; the STATE model is still trained.