

Aligning LLMs with Biomedical Knowledge using Balanced Fine-Tuning

A simple training method for extending domain capability
with less general forgetting

Zhenchao Tang*, Fang Wang*†, Haohuai He, Jiale Zhou, Tianxu Lv, Jun Zhu,
Shouzhi Chen, Minghao Yang, Yu Wang, Jiayang Wu, Yidong Song, Yaokun Li,
Jiehui Huang, Jun Zhou, Bing He, Jianhua Yao†

*Equal contribution †Corresponding authors

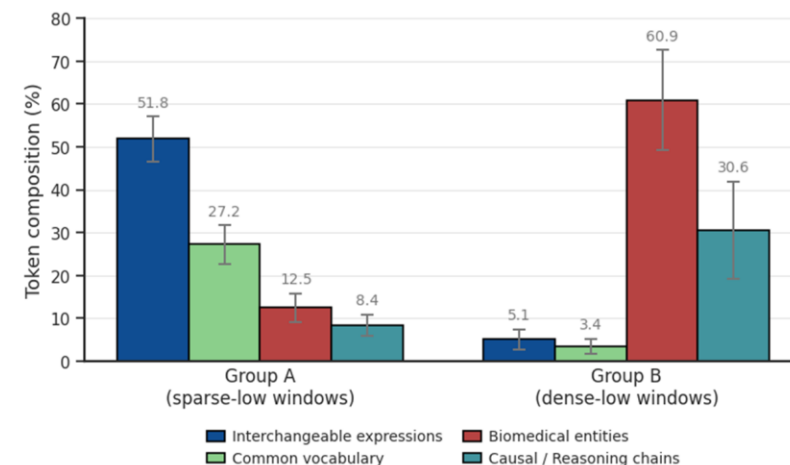
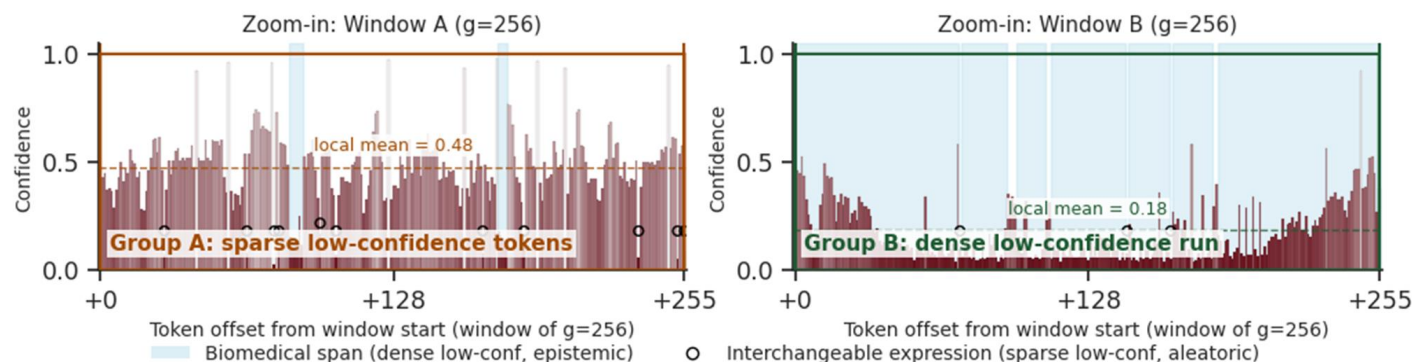
Tencent

 **AI for
Life Sciences Lab**

Paper: arxiv.org/abs/2511.21075

Challenge and motivation: useful difficulty

Low confidence can reflect unfamiliar biomedical knowledge. BFT considers the surrounding passage when deciding what to learn.



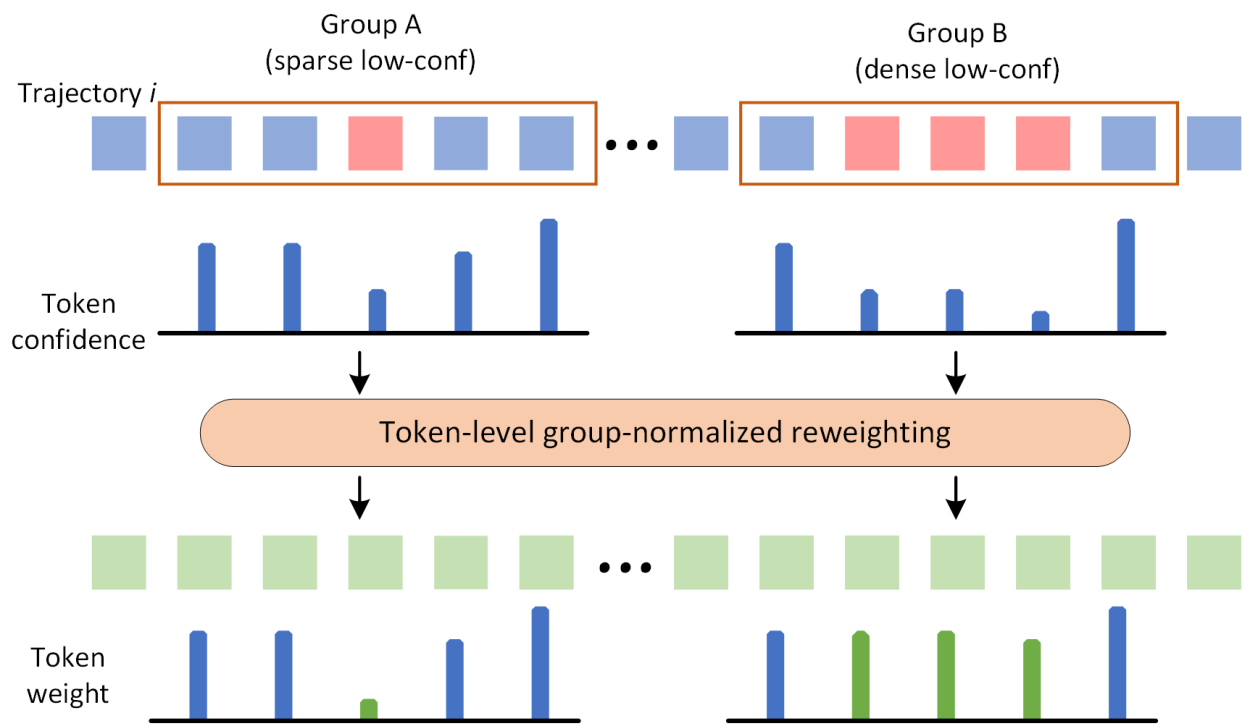
Isolated difficult words and dense difficult passages need different treatment.

Dense difficult regions: 60.9% biomedical entities vs 12.5% in sparse regions.

This association motivates the method; it does not make every difficult word useful.

Method: Balanced Fine-Tuning

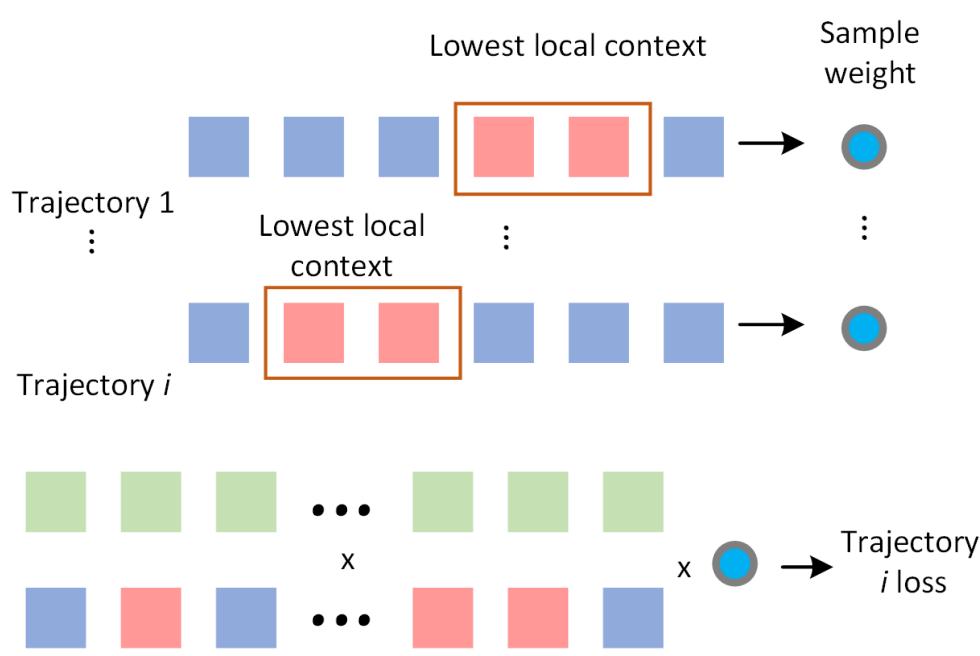
1 Contextual word weighting



Retain learning attention for words inside difficult passages.

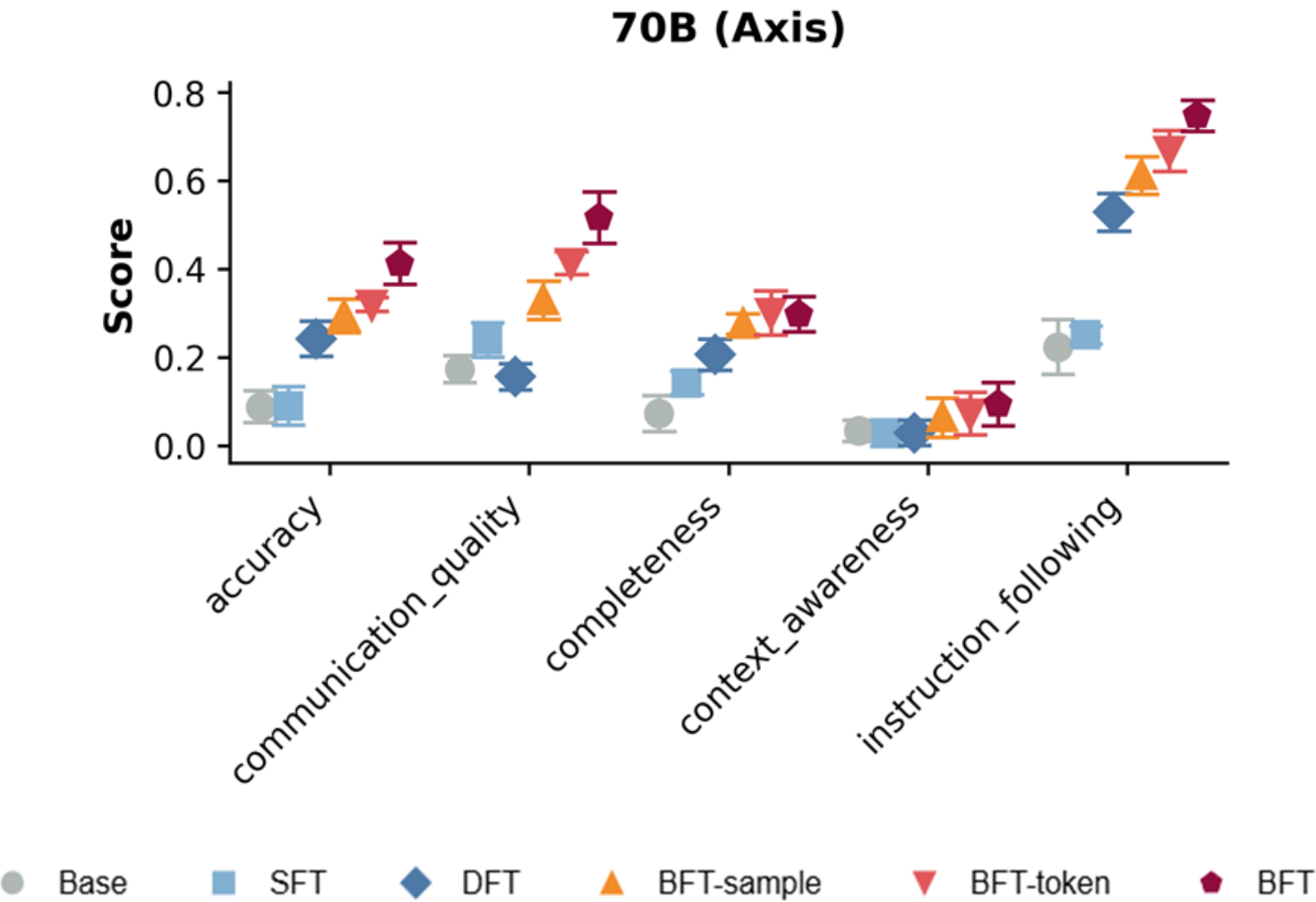
Uses the model's own confidence. No reward model or external uncertainty estimator.

2 Bounded example weighting



Emphasize examples with hard passages and combine both weights.

Results: stronger medical competence

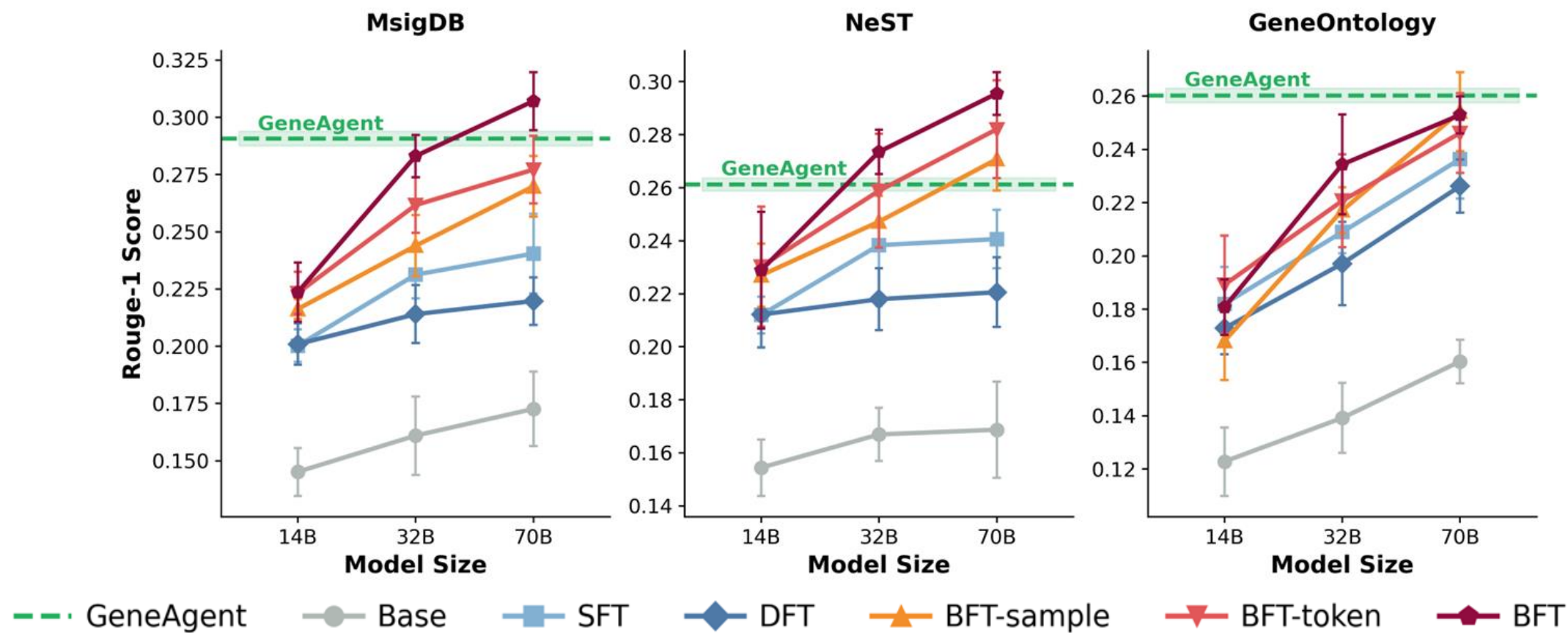


HealthBench Hard · 70B

HealthBench Hard compares five dimensions of medical competence.

SFT, DFT and BFT use matched training recipes. The study evaluates 14B, 32B and 70B models.

Results: better biological process reasoning



BFT achieves the best overall ROUGE among the compared post-trained open-weight models.

Results: causal reasoning and RL compatibility

VCWorld average accuracy · 70B

0.696

BFT before GRPO

0.562

SFT before GRPO

0.736

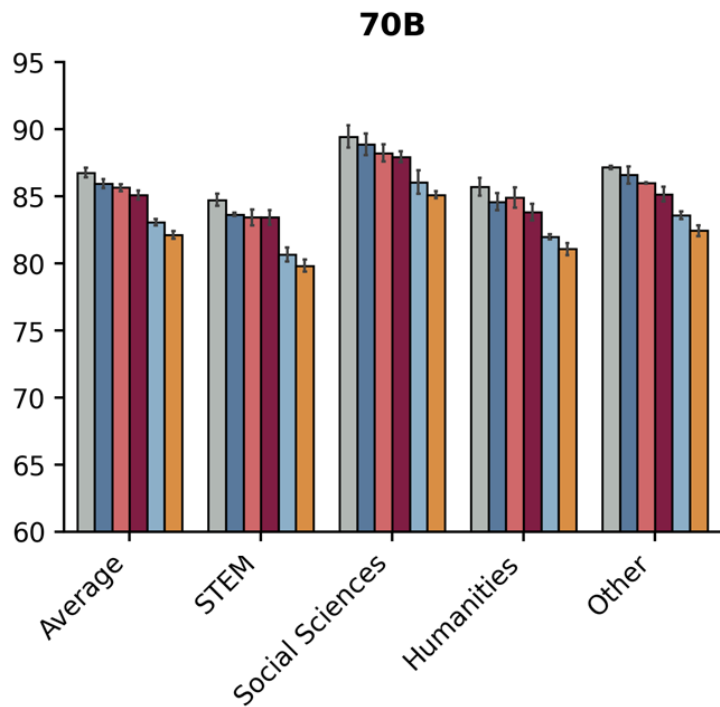
BFT after GRPO

Under the same GRPO setup, BFT improves from 0.696 to 0.736.
SFT declines from 0.562 to 0.486.

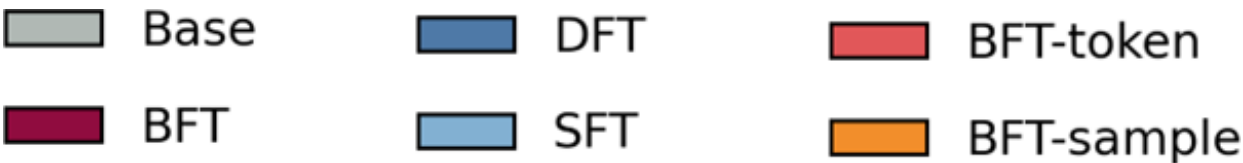
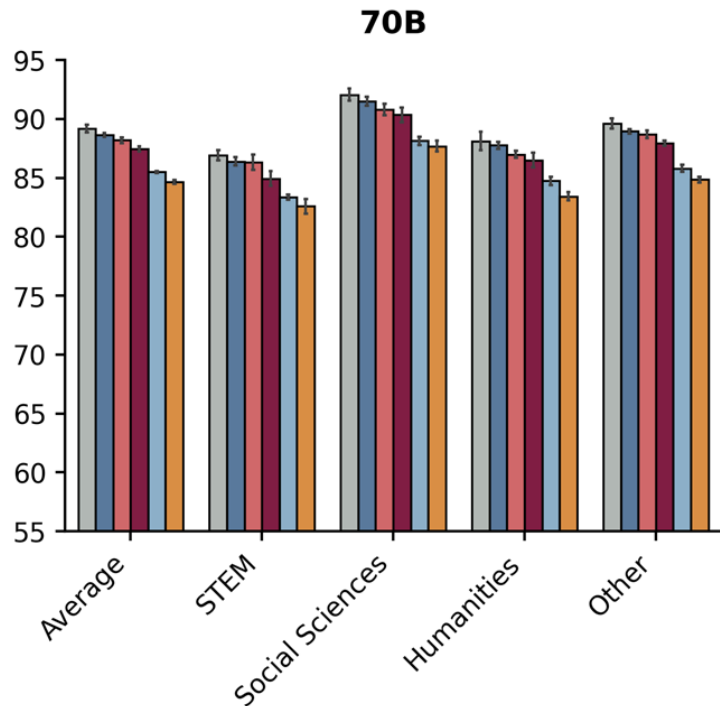
BFT is not statistically distinguishable from Gemini-2.5-Flash across the five held-out cell lines. The RL result is specific to this experimental setting.

Results: less forgetting on general knowledge

MMLU



CMMLU



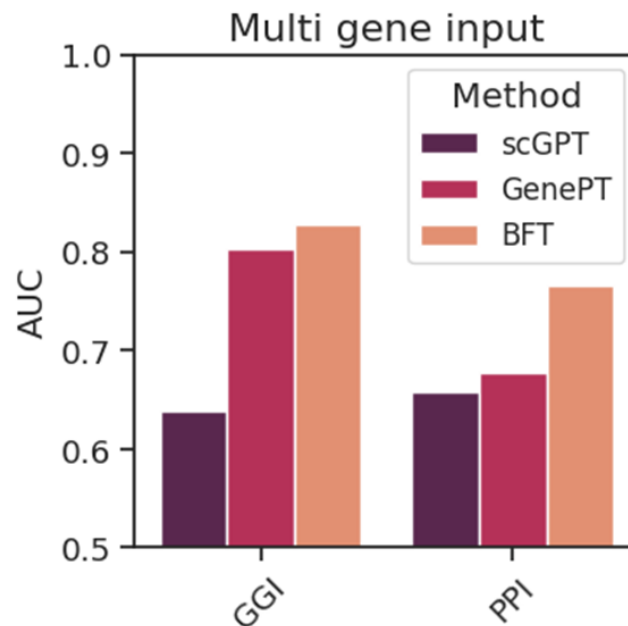
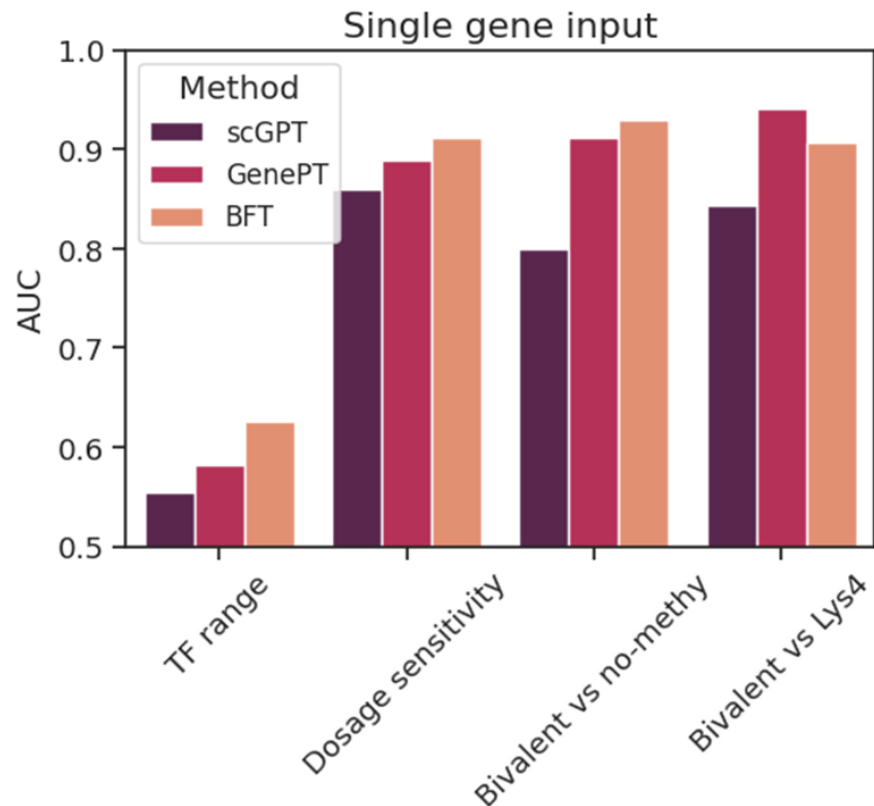
Domain adaptation with retention

BFT preserves more general knowledge than SFT and sample-only BFT.

Some forgetting remains.

DFT forgets least, while BFT offers stronger biomedical adaptation.

Results: useful gene representations



0.829 vs 0.802

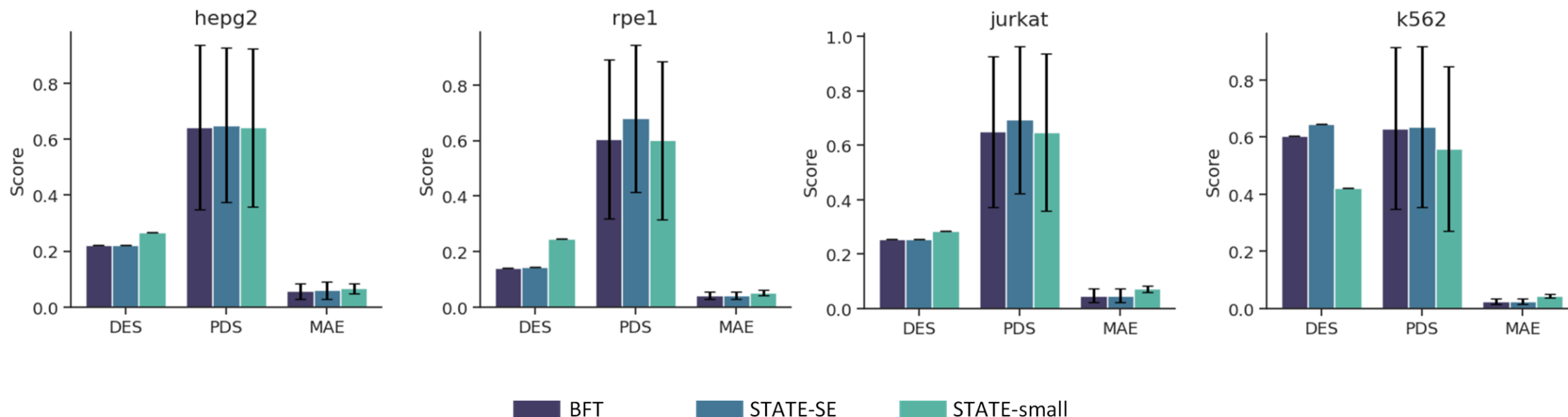
Mean gene score
BFT vs GenePT

Both use the same text
encoder, with different
profile sources.

GenePT remains higher on
Lys4.

BFT-generated profiles support downstream prediction of gene properties.

Results: single-cell perturbation prediction



BFT embeddings perform comparably to standard STATE inputs across four perturbation datasets.

Profile generation is zero-shot with respect to these datasets. The STATE decoder is still trained.

BFT: expanding domain capability with less forgetting

A simple route toward extending LLMs into STEM

Contextual word weighting preserves learning signals in difficult passages.
Example weighting emphasizes passages that need more learning.

The demonstrated gains are in medicine and biology. Broader STEM transfer remains a direction for future evaluation.

Paper arxiv.org/abs/2511.21075

Code github.com/TencentAILabHealthcare/BFT

Project shapsider.github.io/BFT/