Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models

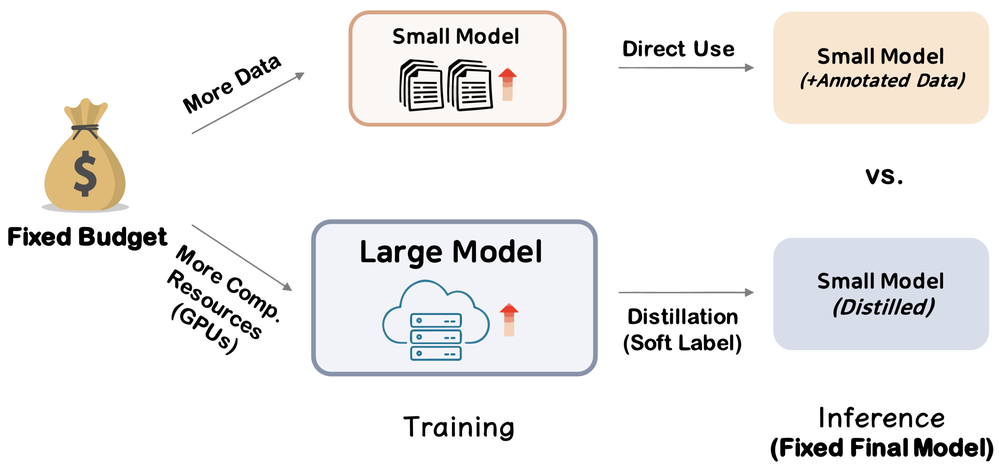
Junmo Kang, Wei Xu, Alan Ritter

Junmo.kang@gatech.edu; {wei.xu, alan.ritter}@cc.gatech.edu

Georgia Tech Machine Learning

Research Question

Q. Given a fixed budget, how to build a compact model in a cost-efficient way?



Trade-Off

Annotation Cost

Strategy 1: annotate more data to directly train a small model

vs.

Strategy 2: train a larger model, then distill into a small model

Computational Cost

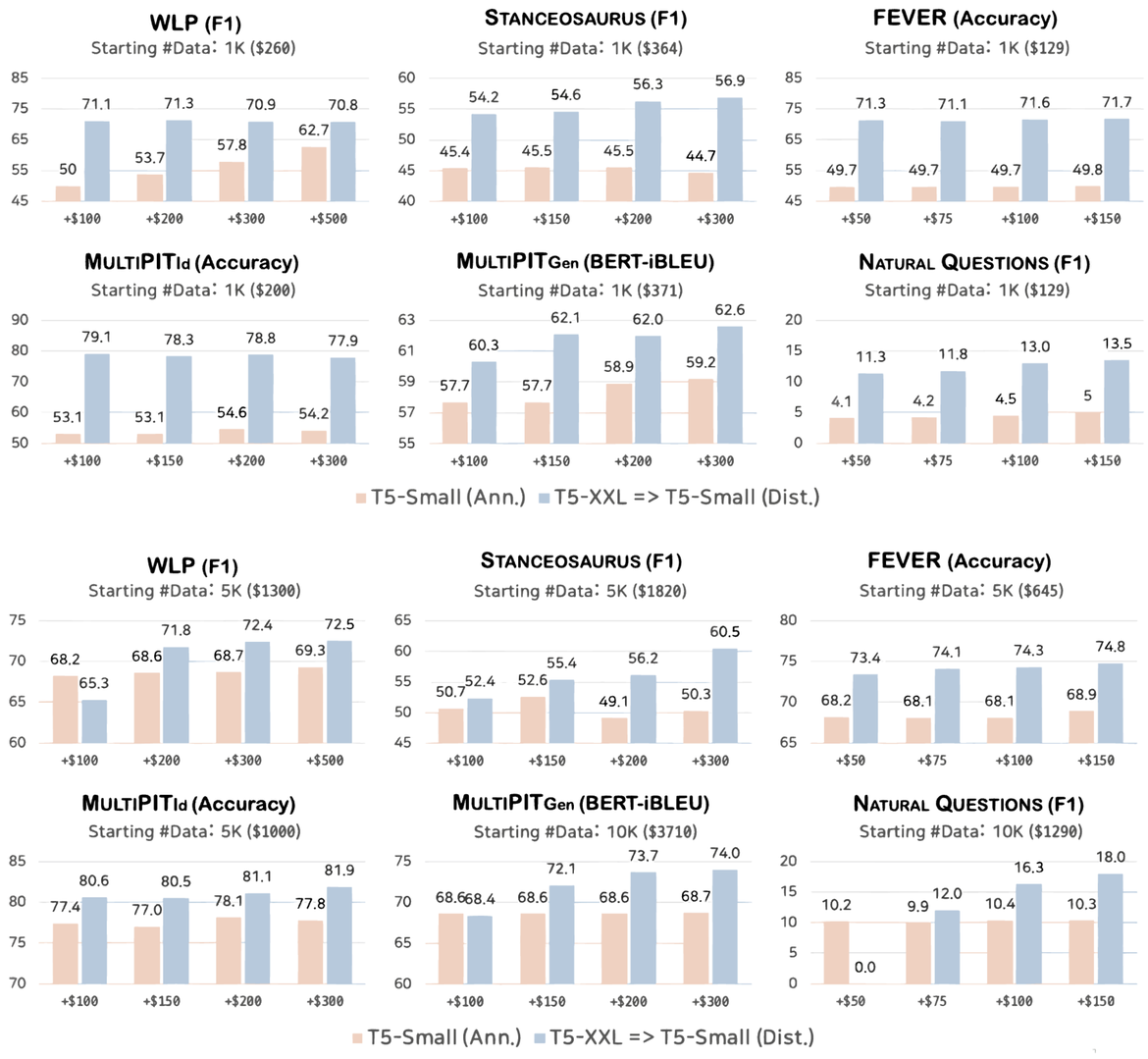
Task & Annotation Cost

Dataset	Task	$ per Label
WLP	Named Entity Recognition	$0.260
STANCEOSAURUS	Stance Classification	$0.364
FEVER	Fact Verification	$0.129
MULTIPITID	Paraphrase Identification	$0.200
MULTIPITGEN	Paraphrase Generation	$0.371
Natural Questions	Question Answering	$0.129

Computational Cost

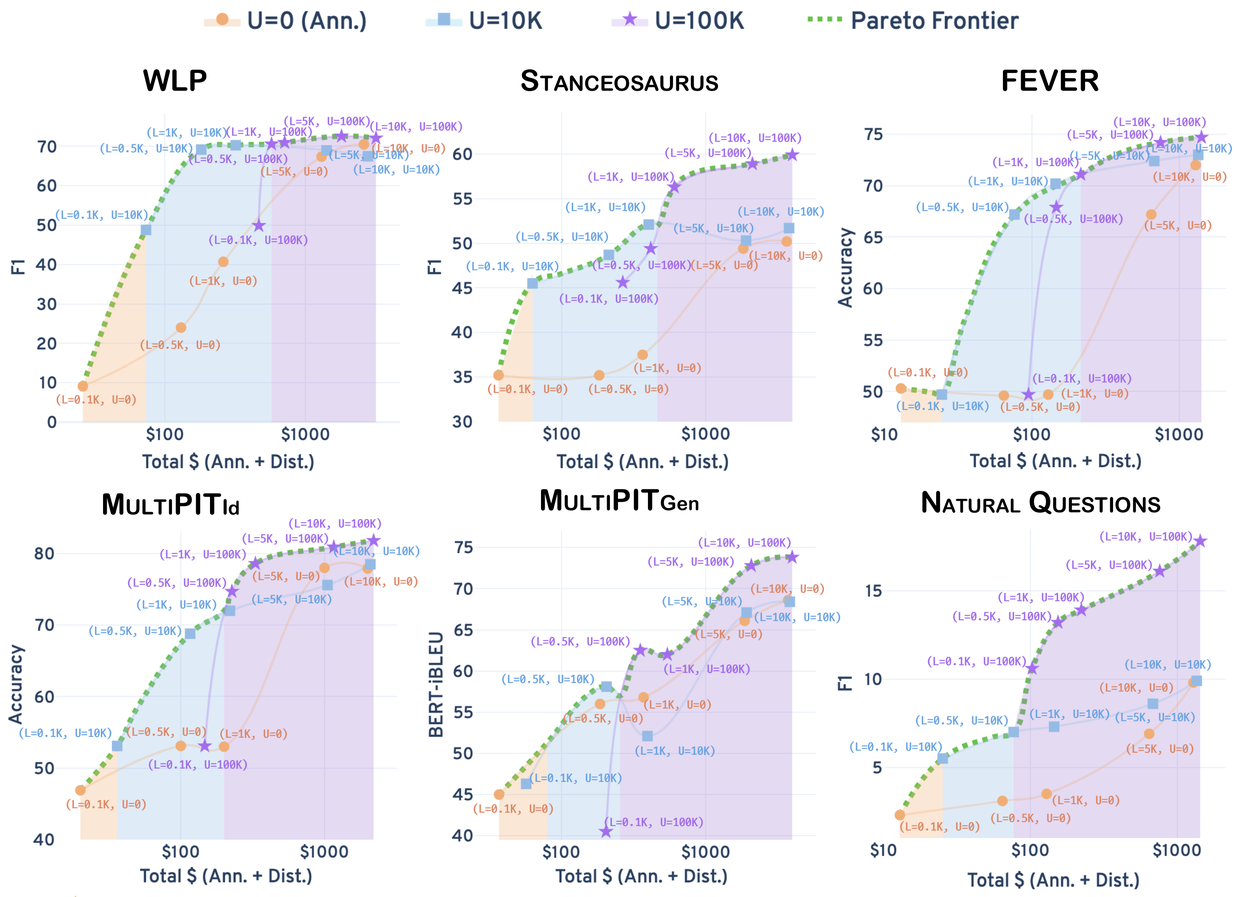
$1.875 per 1 GPU hour (est. based on A100 in Google Cloud Platform)

Main Results



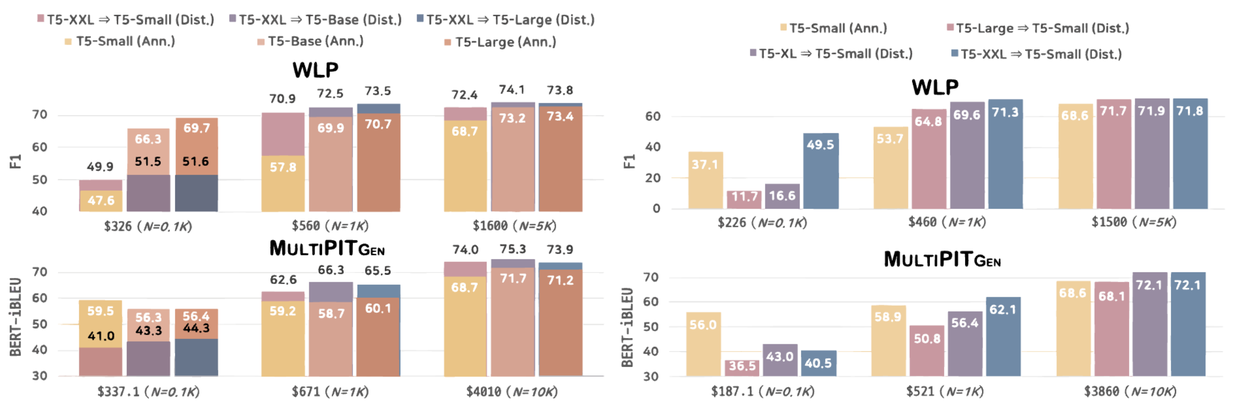
Given fixed budgets, the distillation strategy is more economically efficient

Pareto Curve



Surprisingly, the distillation strategy is Pareto optimal across almost all budgets

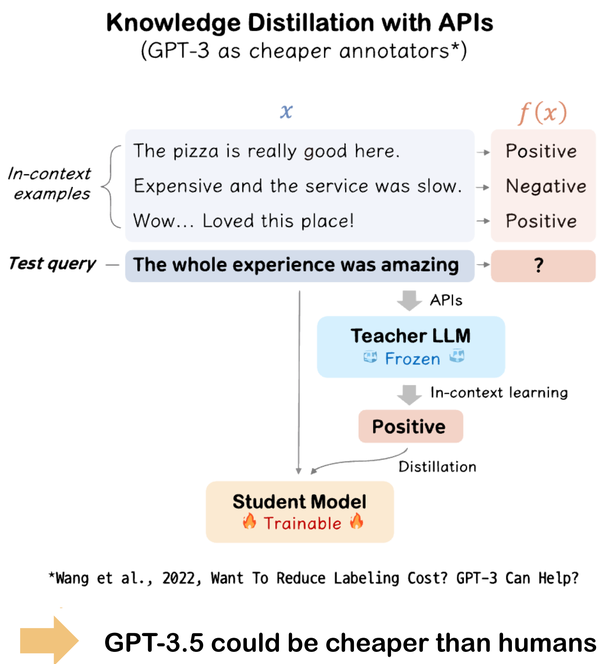
Analysis with Different Small & Large Models

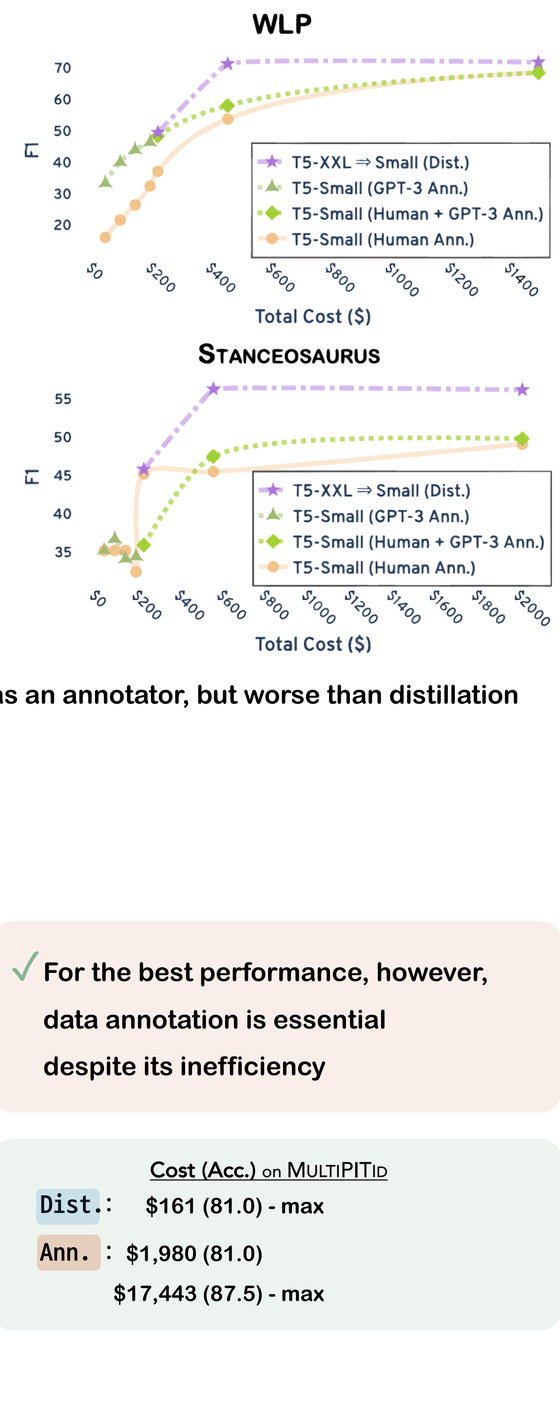


A smaller final model could be better in both performance and inference efficiency

The largest teacher model is not necessarily the best

GPT-3.5 as an Annotator





GPT-3.5 could be cheaper than humans as an annotator, but worse than distillation

Takeaways

✓ In general, data annotation might not be the best practical solution in light of cost-efficiency; Scale up, then distill !

✓ For the best performance, however, data annotation is essential despite its inefficiency

✓ Synthetic data generation using GPT-3.5 could be cost-efficient compared to humans, but still limited

Cost (Acc.) on MULTIPITID

Dist.: $161 (81.0) - max

Ann.: $1,980 (81.0)

$17,443 (87.5) - max