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

Junmo Kang

Wei Xu

Alan Ritter

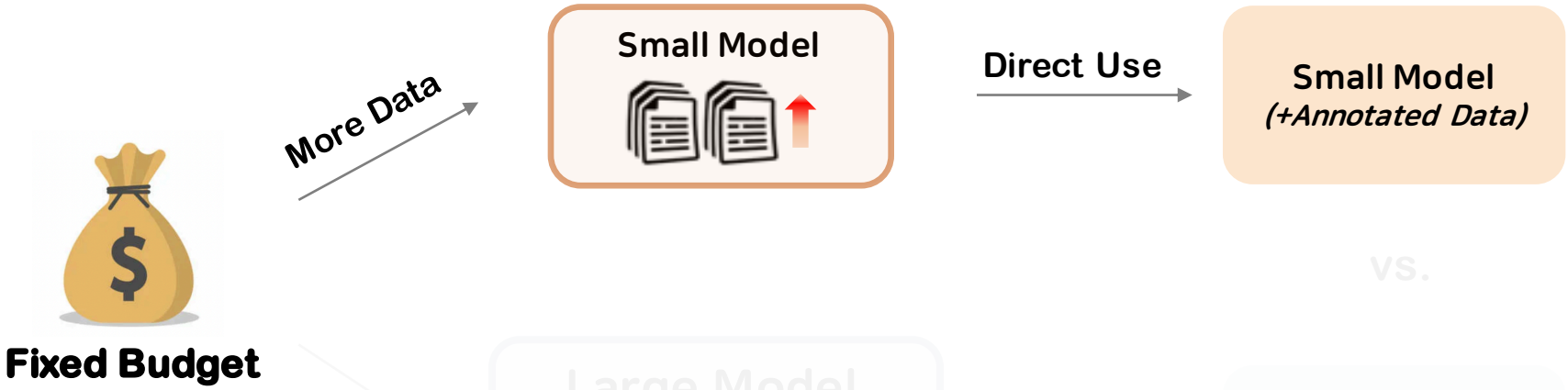
Georgia Institute of Technology

61 ACL 2023

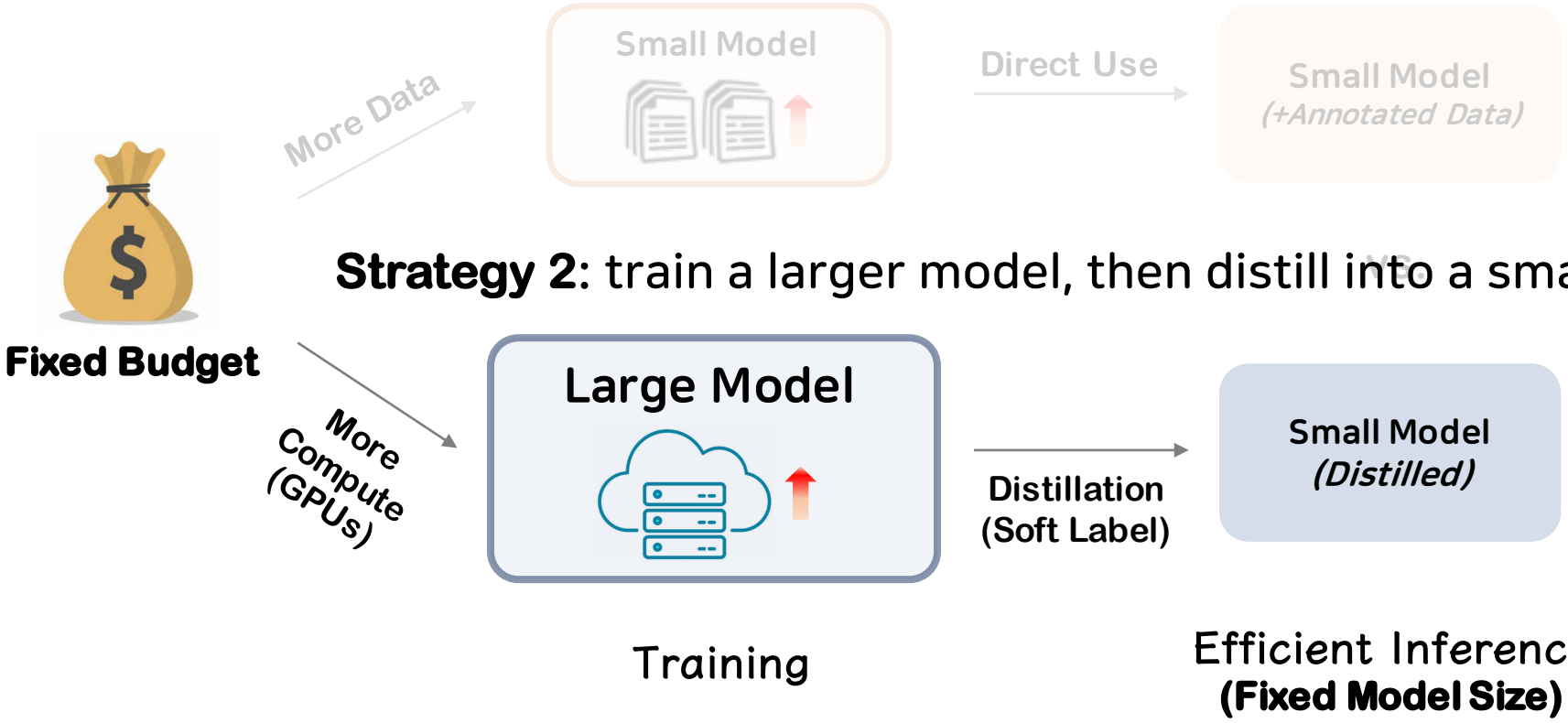
Q. Given a fixed budget,
how to build a compact model in a cost-efficient way?

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

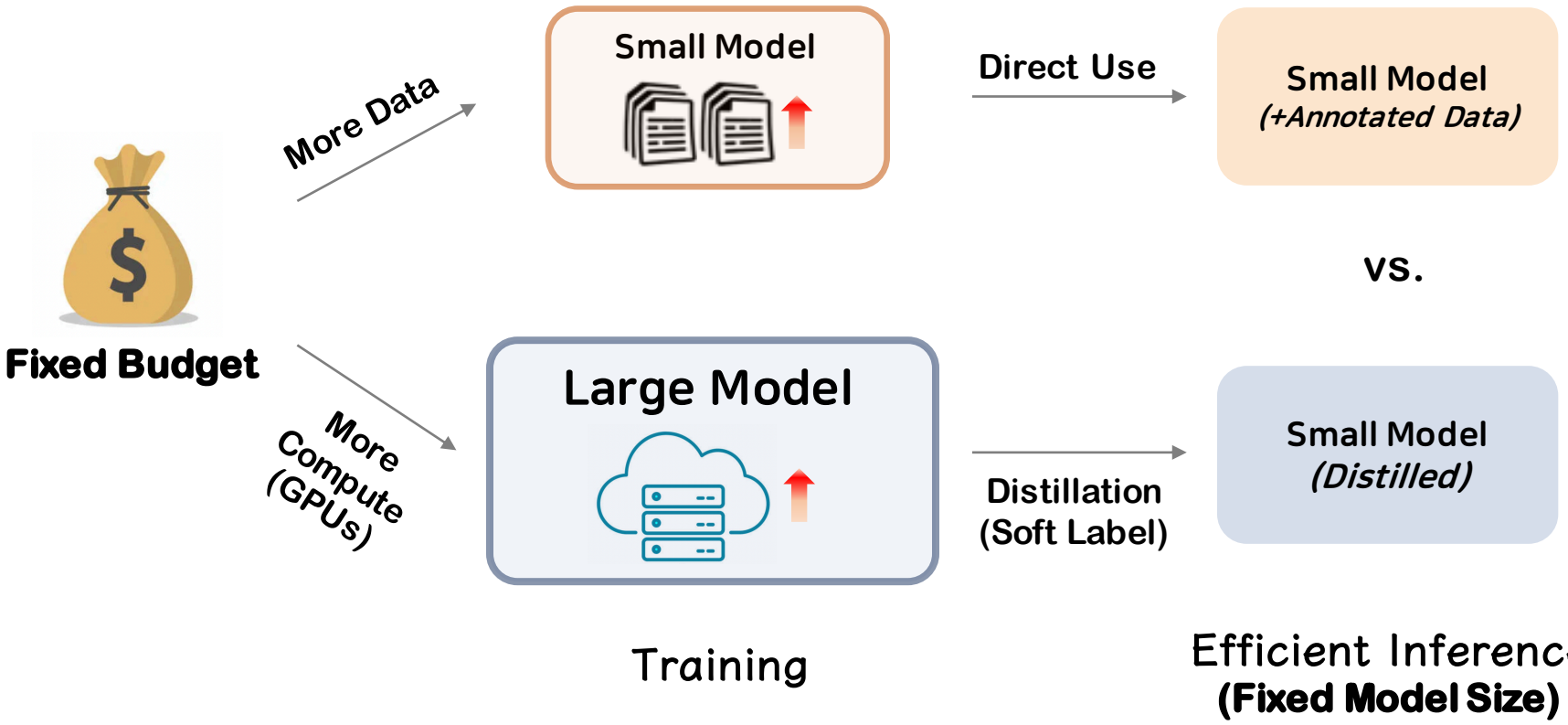
Strategy 1: annotate more data to directly train a small model



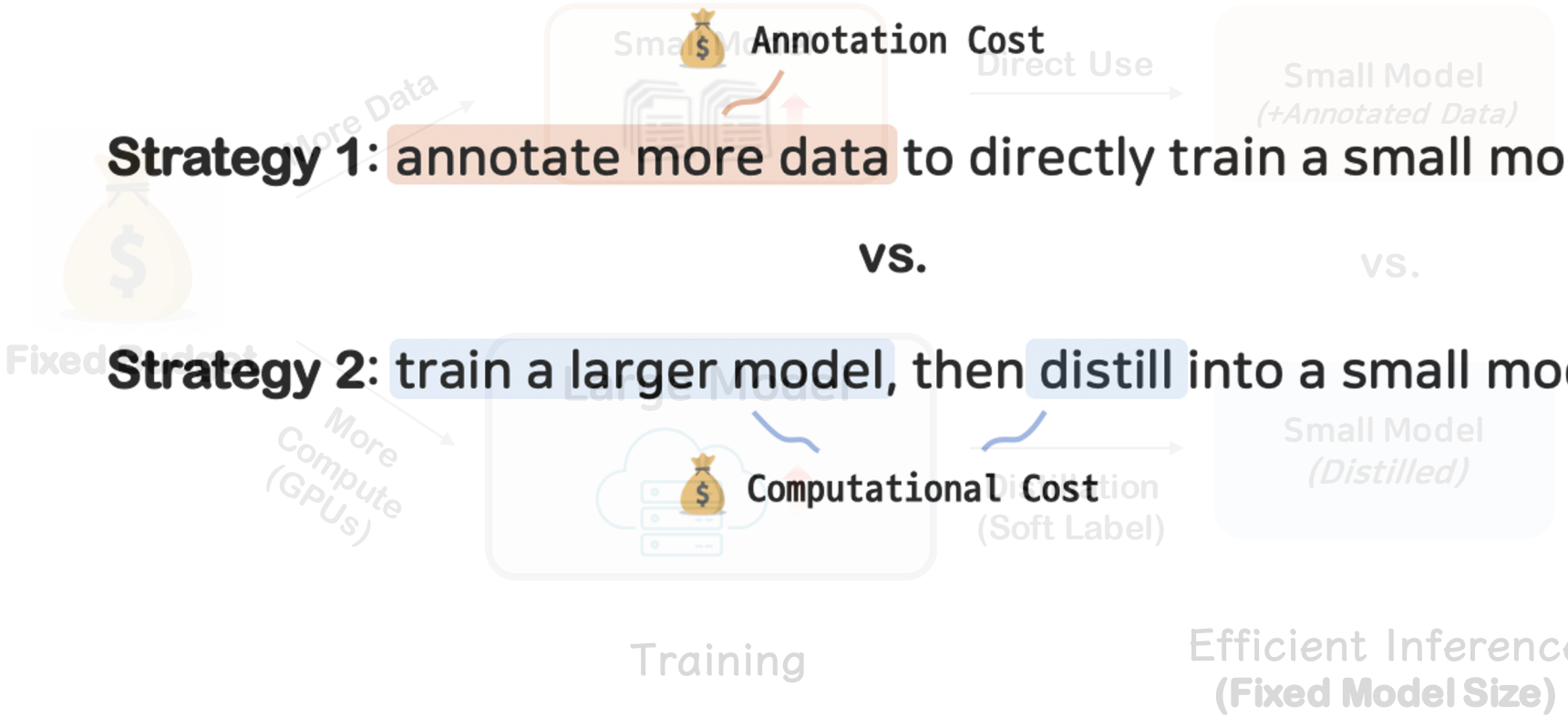
Q. Given a fixed budget, how to build a compact model in a cost-efficient way?



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
Training
Efficient Inference
(Fixed Model Size)

Cost Estimations

Task & Annotation Cost

Dataset	Task	$ per Label
WLP	Named Entity Recognition	$0.26
STANCEOSAURUS	Stance Classification	$0.364
FEVER	Fact Verification	$0.129
MULTIPITID	Paraphrase Identification	$0.2
MULTIPITGEN	Paraphrase Generation	$0.371
Natural Questions	Question Answering	$0.129

Cost Estimations

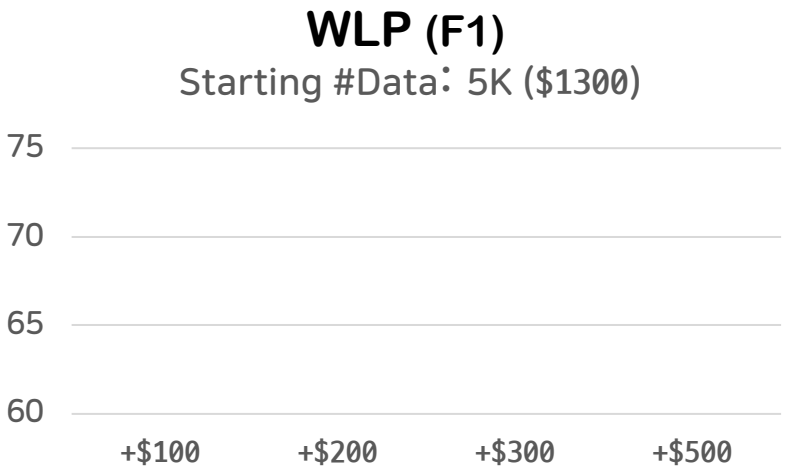
Task & Annotation Cost

Dataset	Task	$ per Label
WLP	Named Entity Recognition	$0.26
STANCEOSAURUS	Stance Classification	$0.364
FEVER	Fact Verification	$0.129
MULTIPITID	Paraphrase Identification	$0.2
MULTIPITGEN	Paraphrase Generation	$0.371
Natural Questions	Question Answering	$0.129

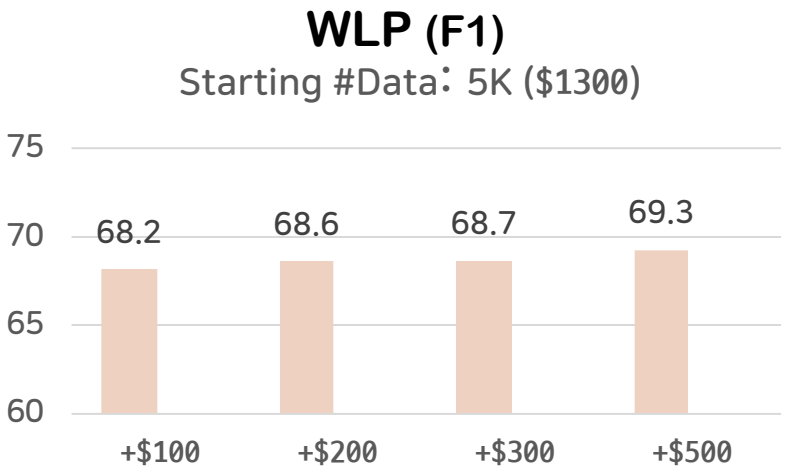
Computational Cost

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

Main Results



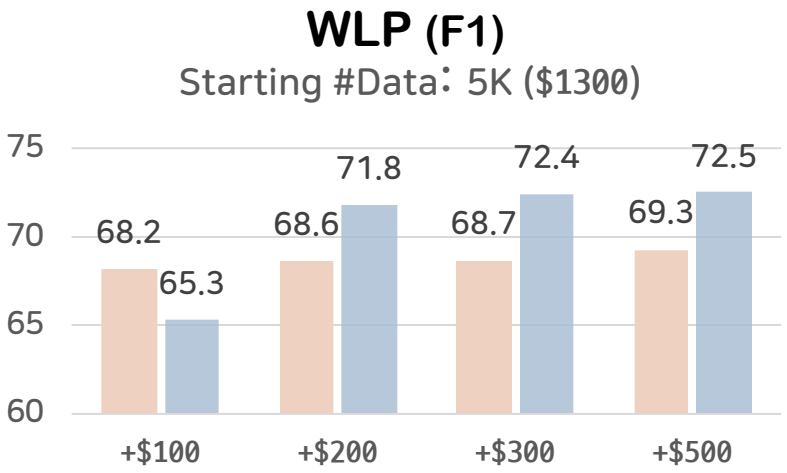
Main Results



Annotation Cost

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

Main Results



💰 Annotation Cost

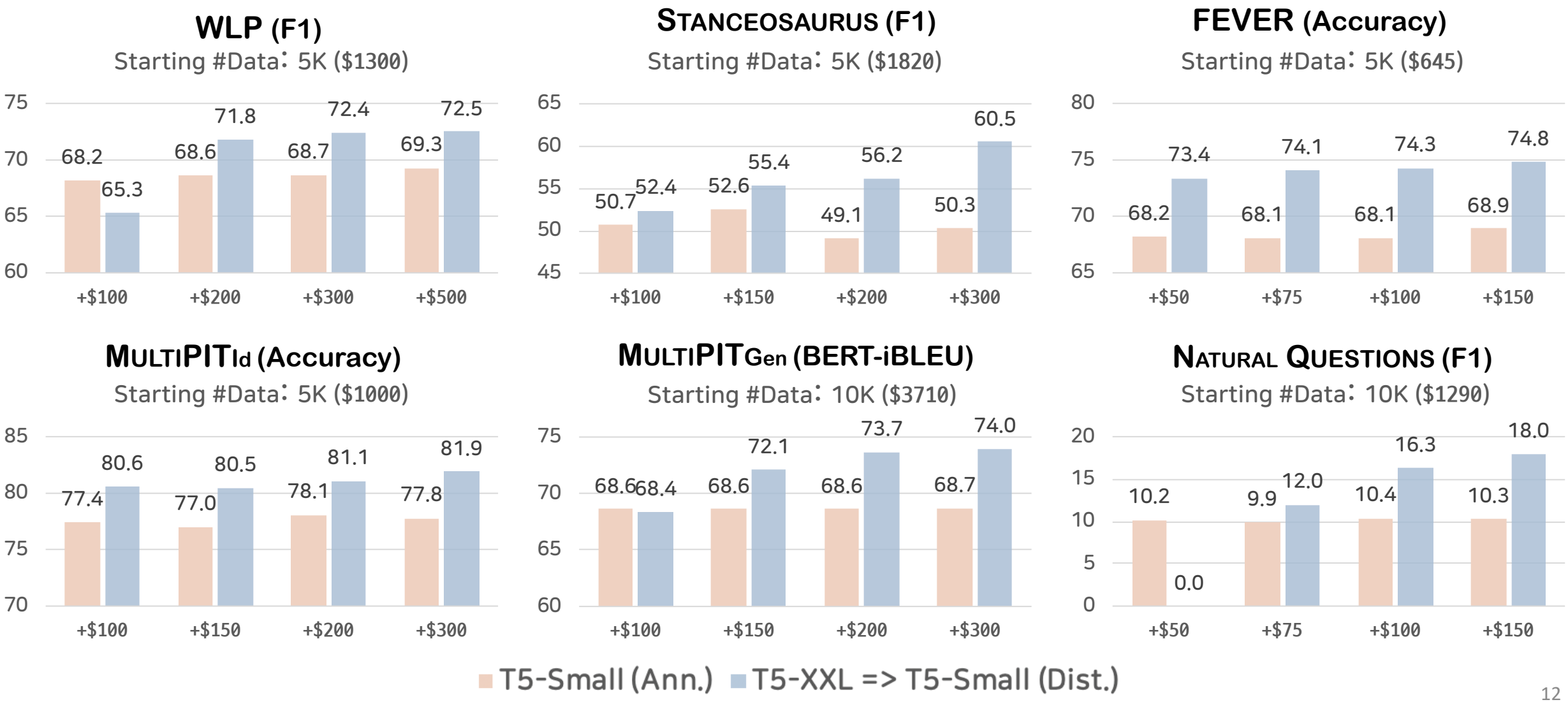
Strategy 1: annotate more data to directly train a small model

vs.

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

💰 Computational Cost

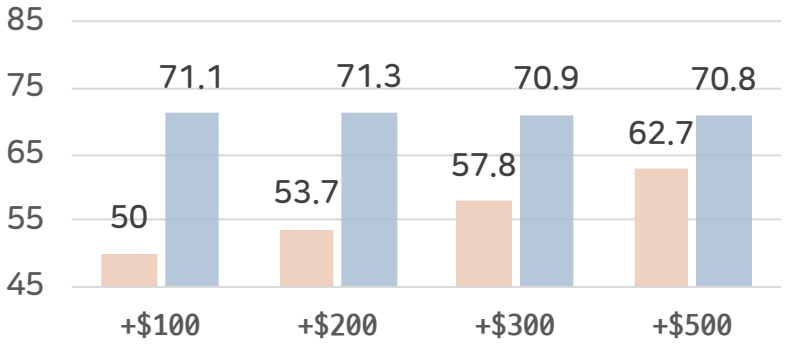
Main Results



Main Results

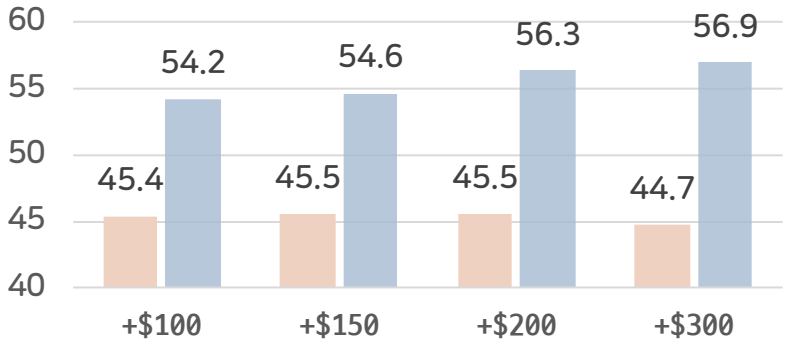
WLP (F1)

Starting #Data: 1K ($260)



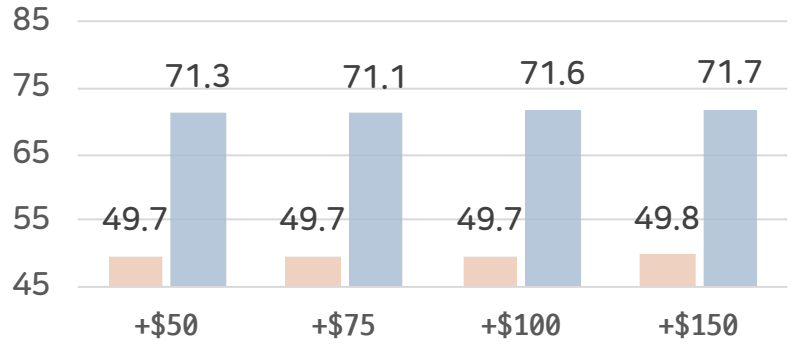
StanceOsaurus (F1)

Starting #Data: 1K ($364)



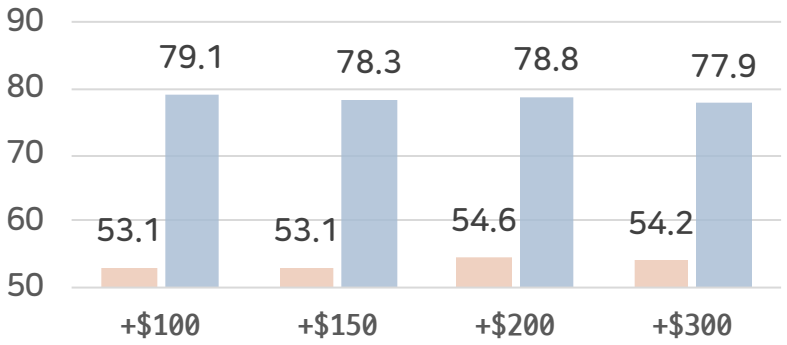
FEVER (Accuracy)

Starting #Data: 1K ($129)



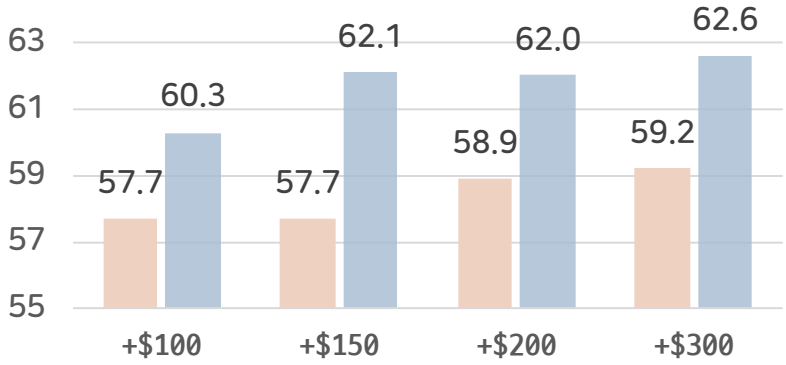
MultiPIT$_{Id}$ (Accuracy)

Starting #Data: 1K ($200)



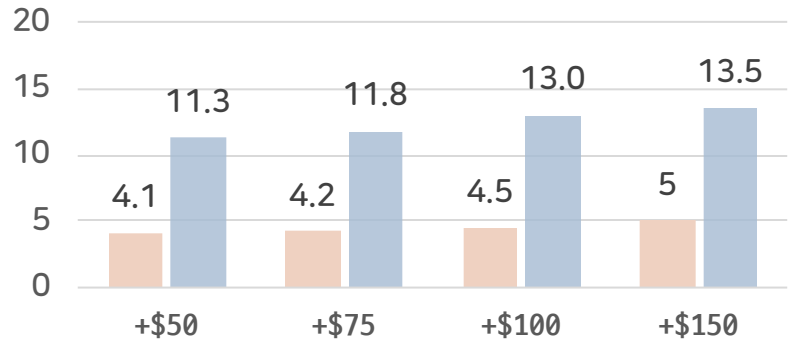
MultiPIT$_{Gen}$ (BERT-iBLEU)

Starting #Data: 1K ($371)



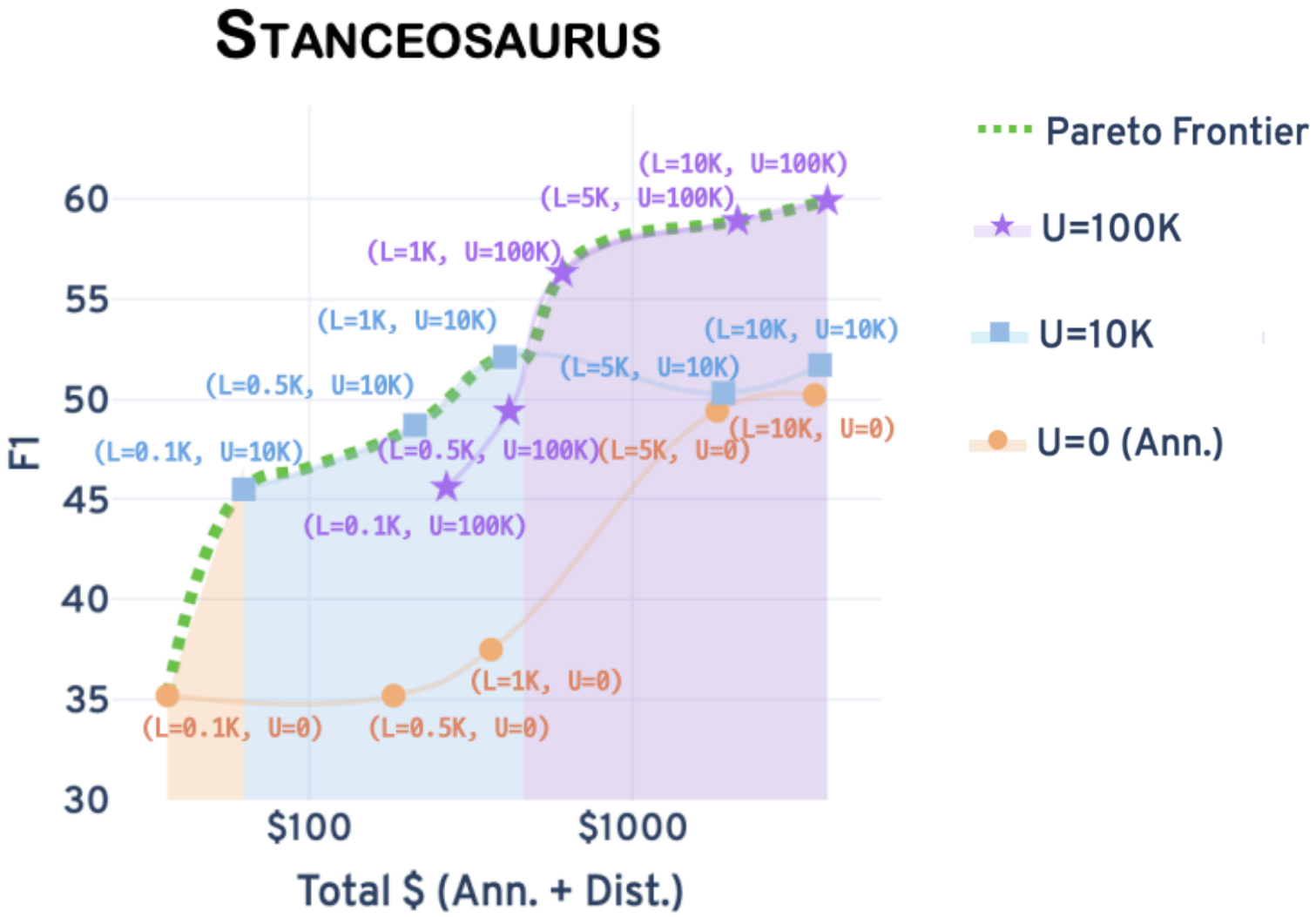
Natural Questions (F1)

Starting #Data: 1K ($129)

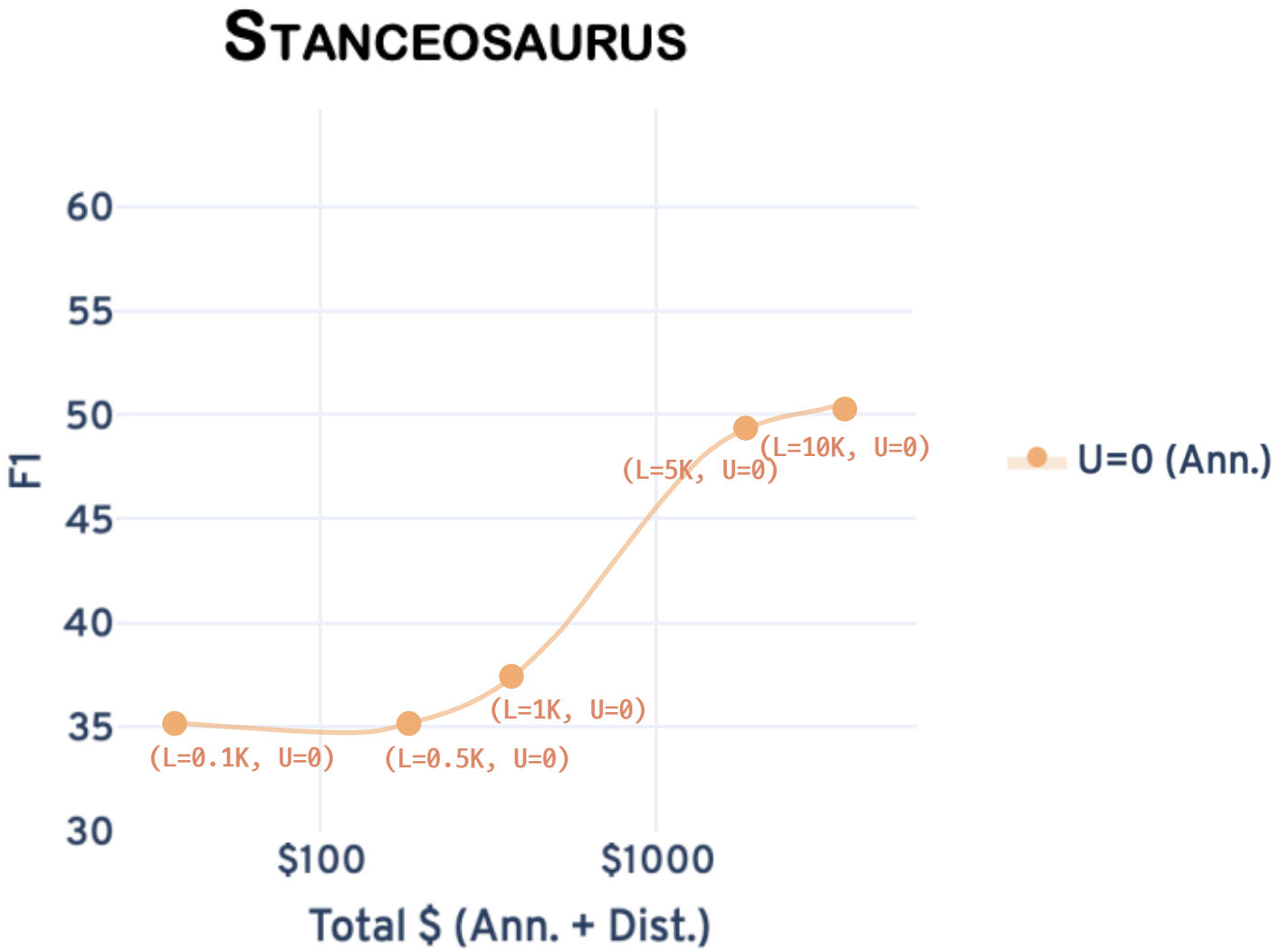


T5-Small (Ann.) T5-XXL => T5-Small (Dist.)

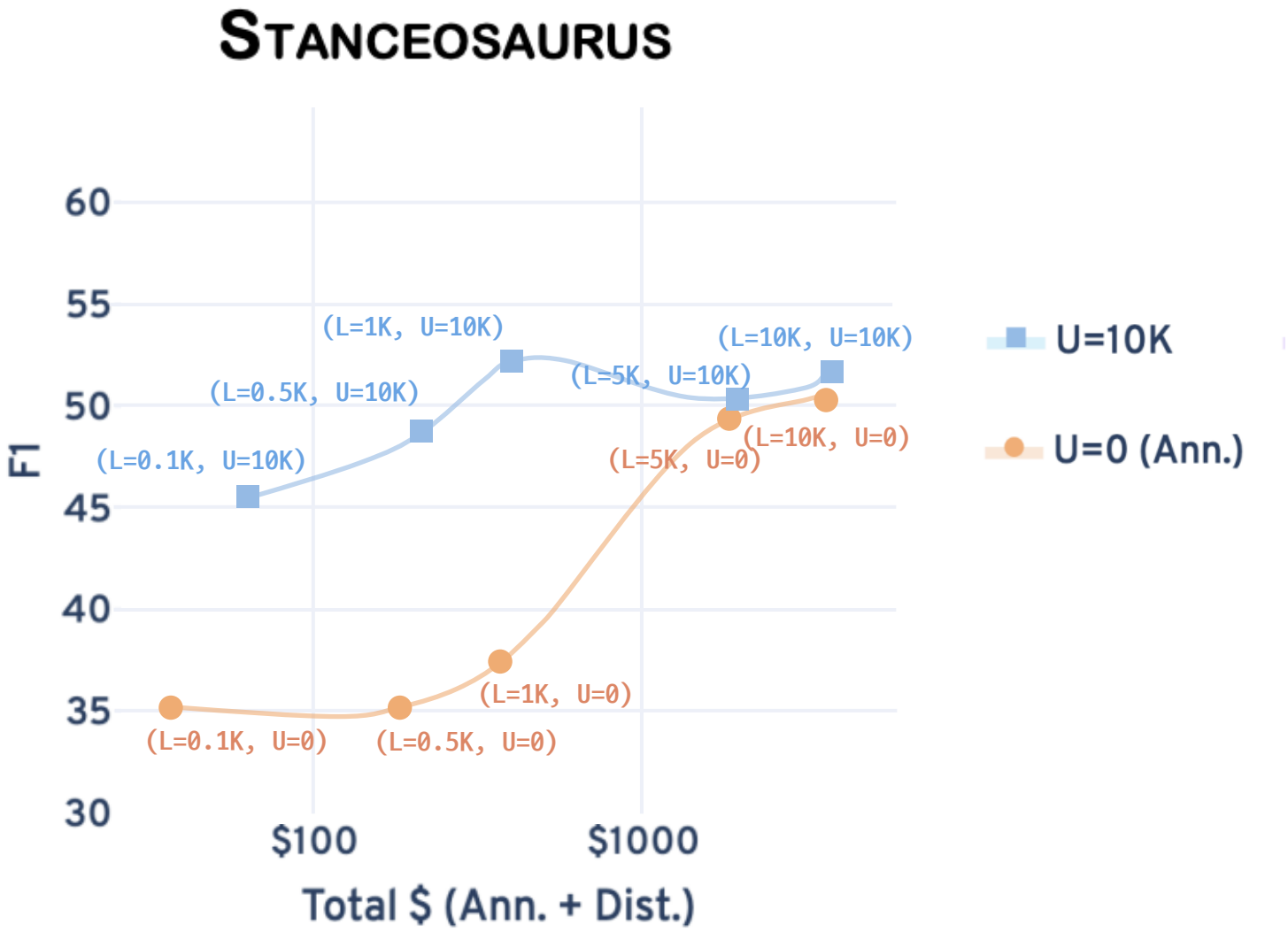
Pareto Curves



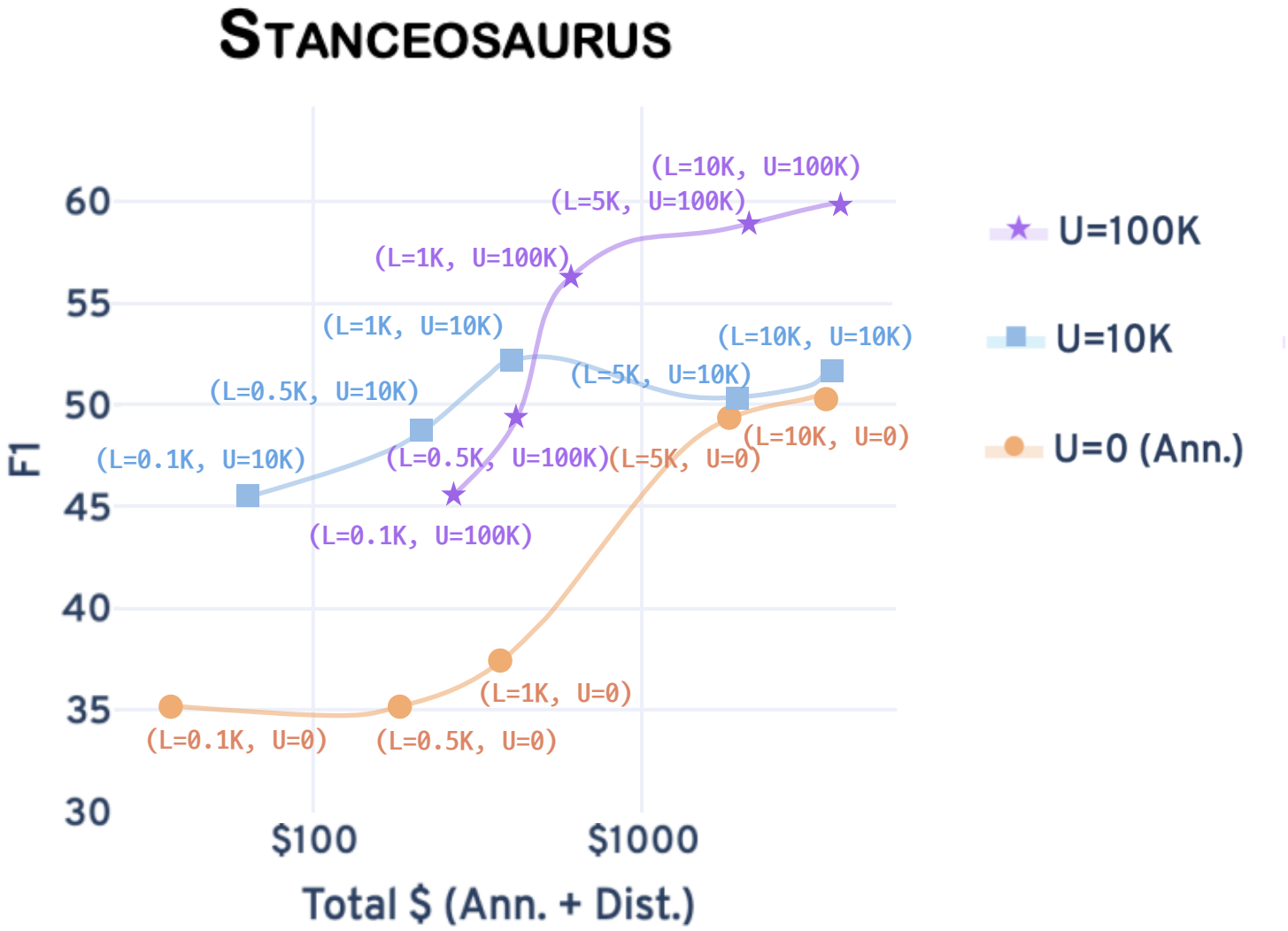
Pareto Curves



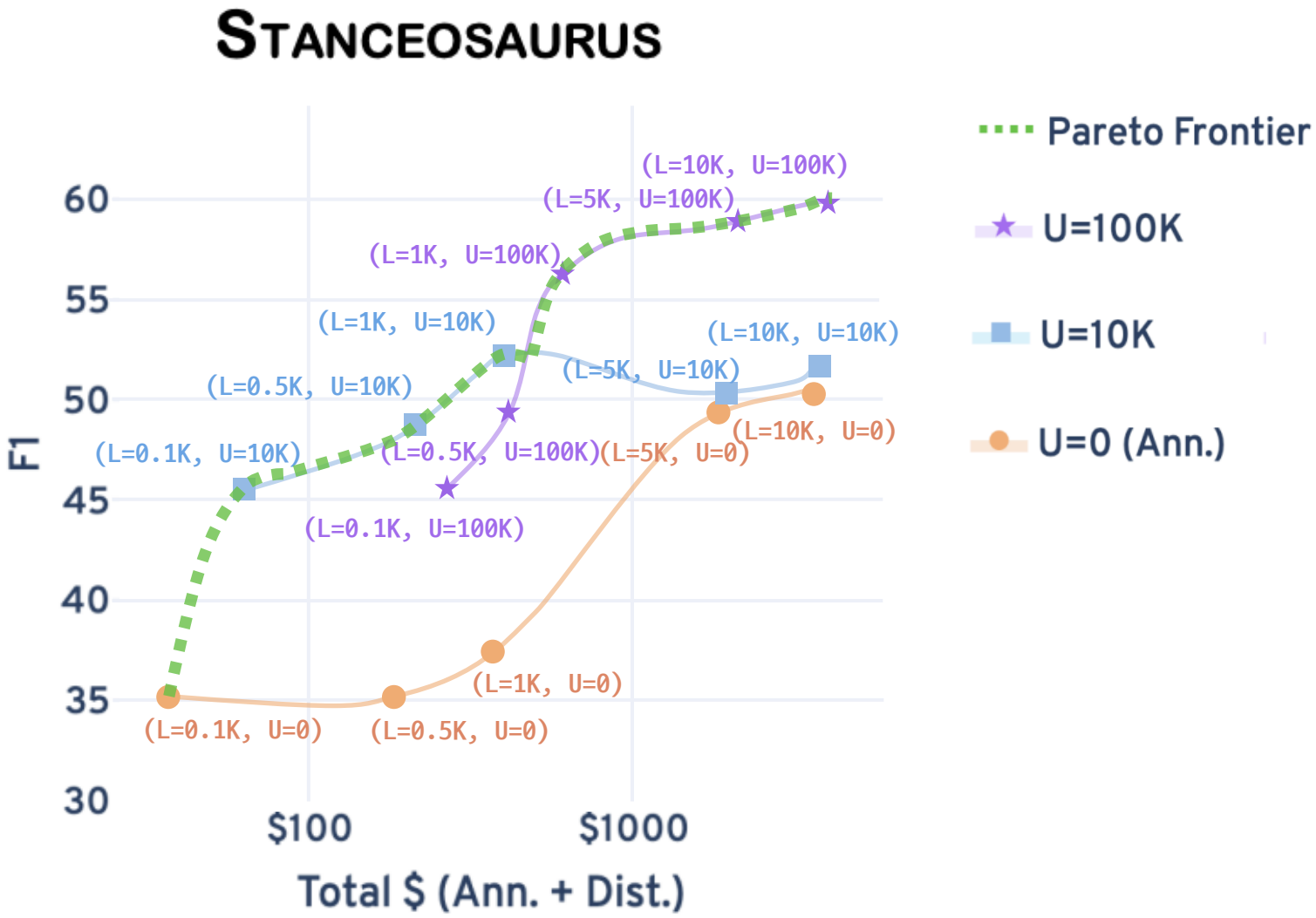
Pareto Curves



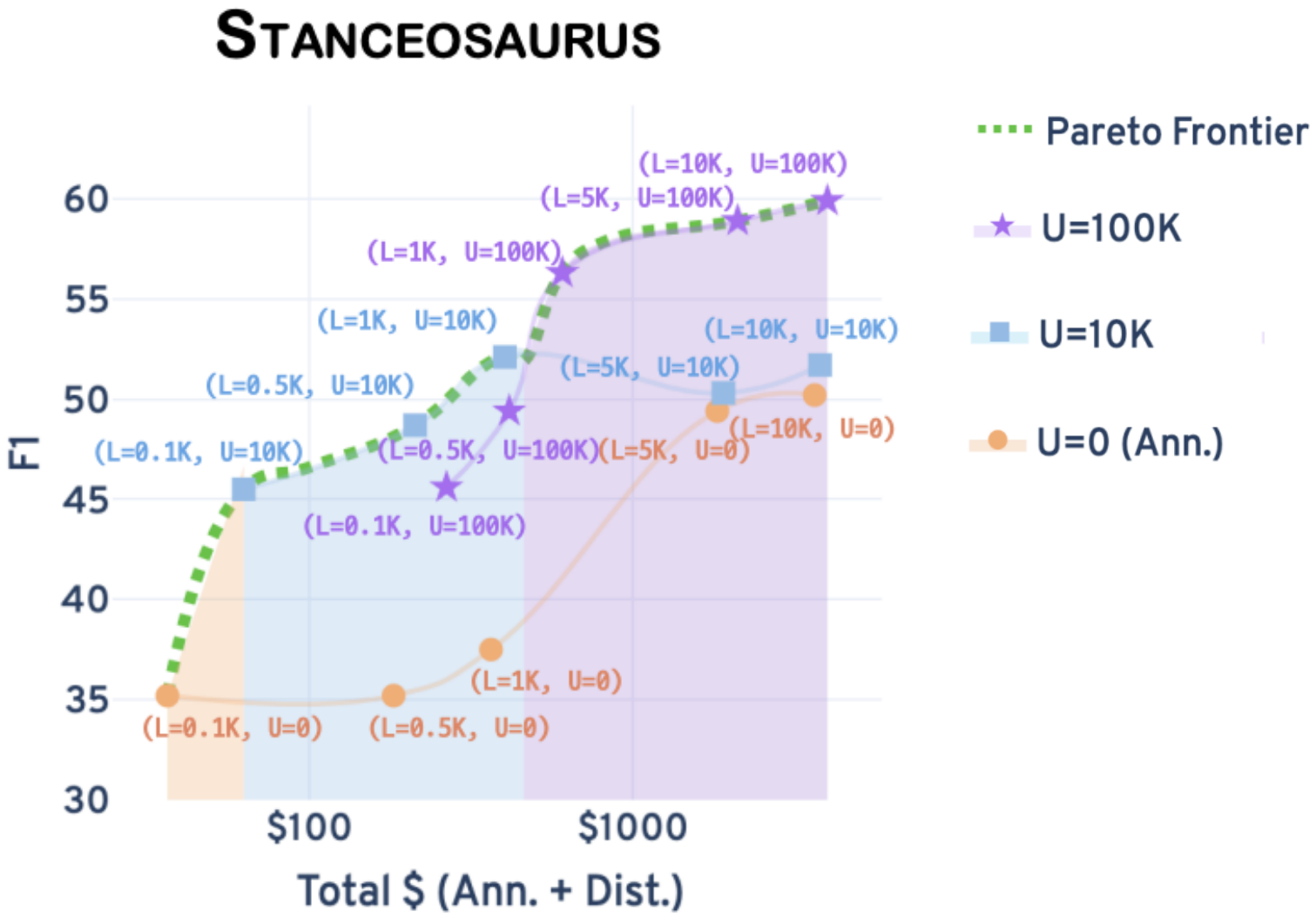
Pareto Curves



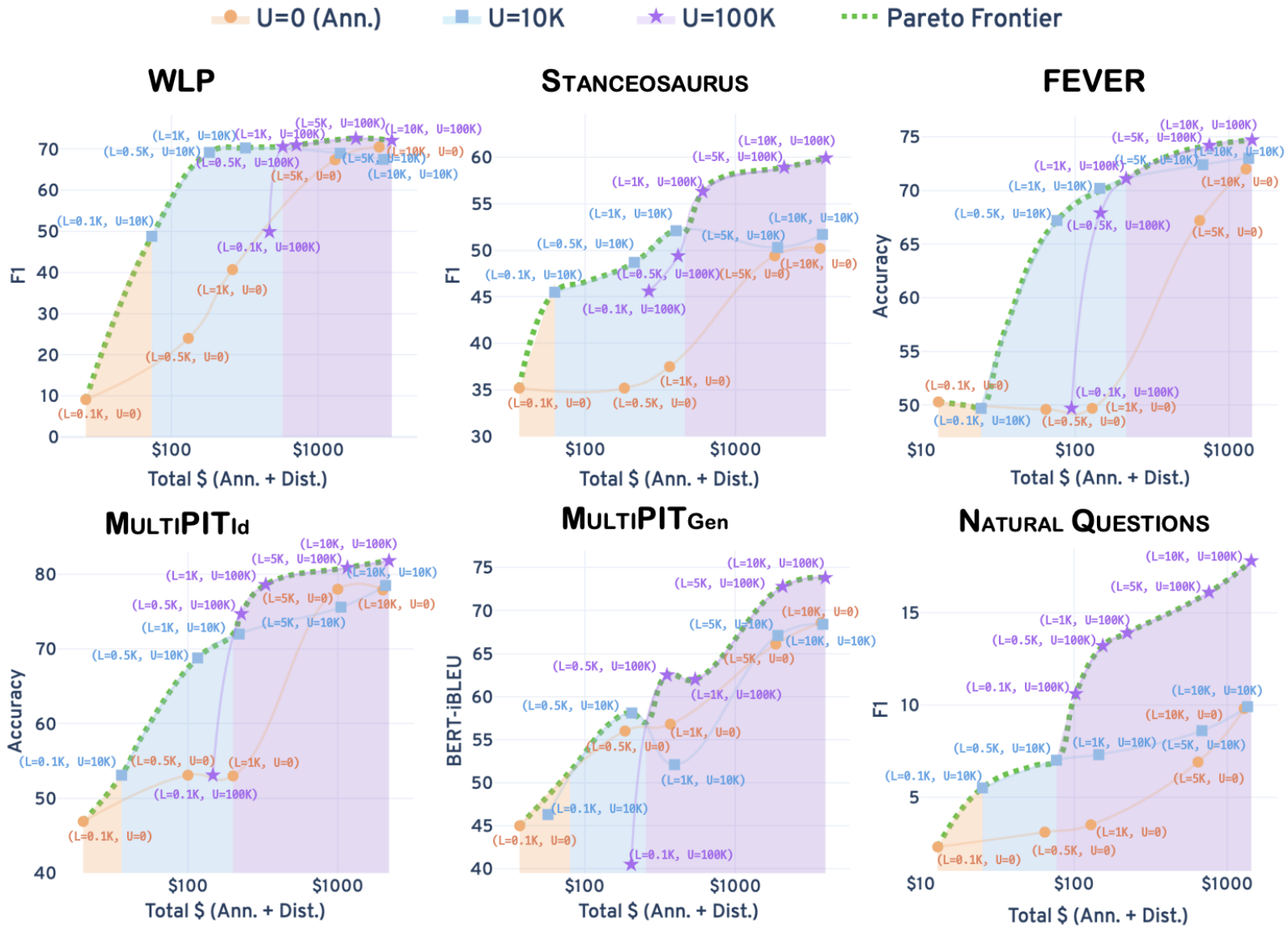
Pareto Curves



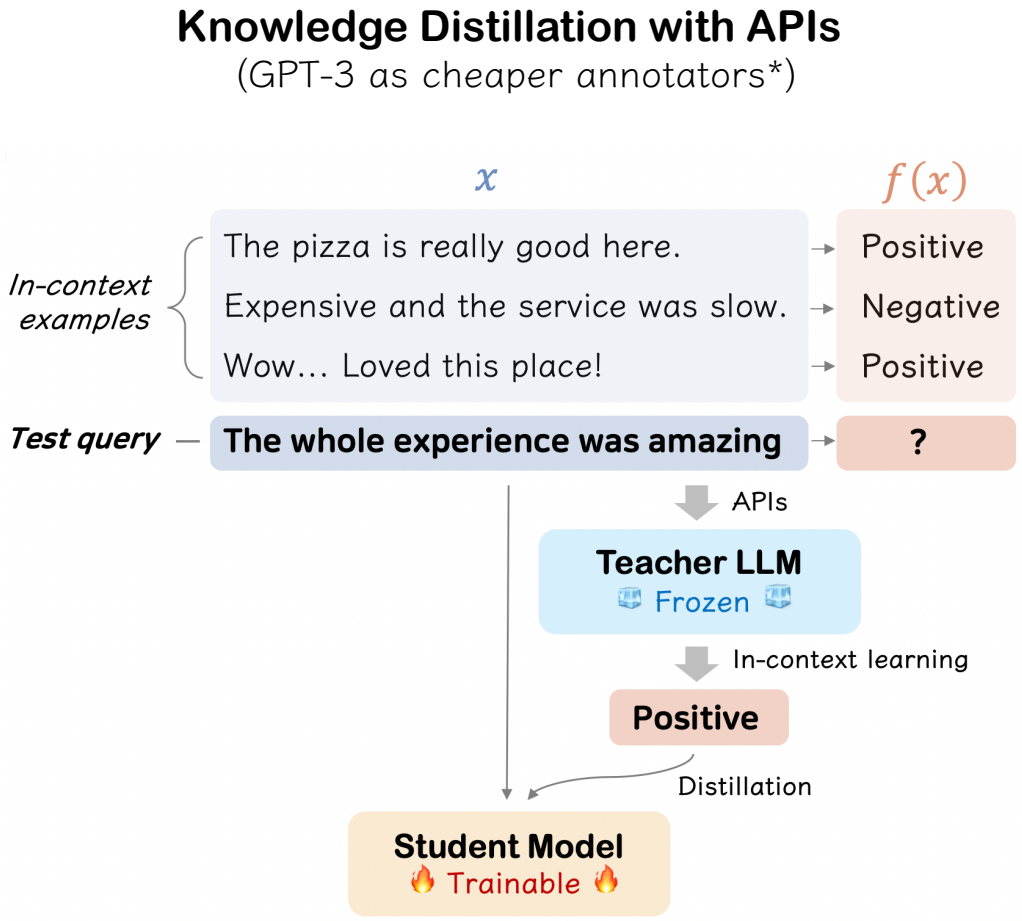
Pareto Curves



Pareto Curves

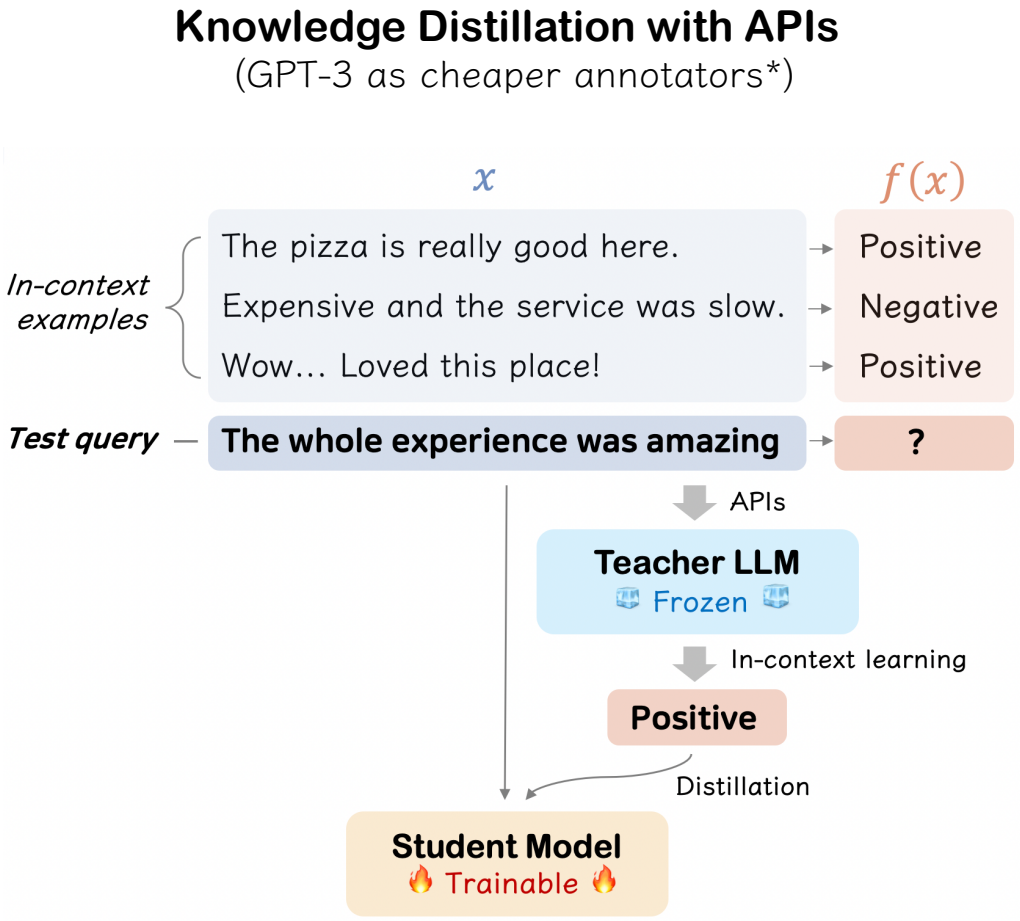


GPT-3.5 as an Annotator

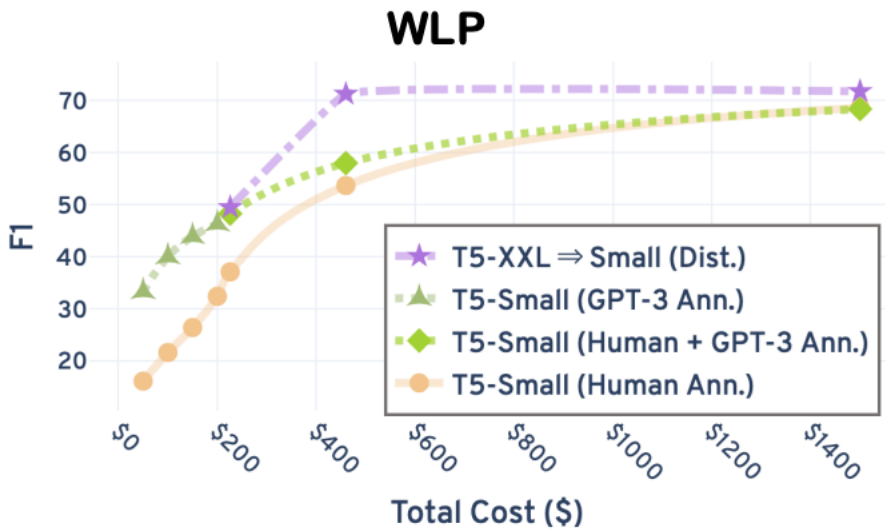


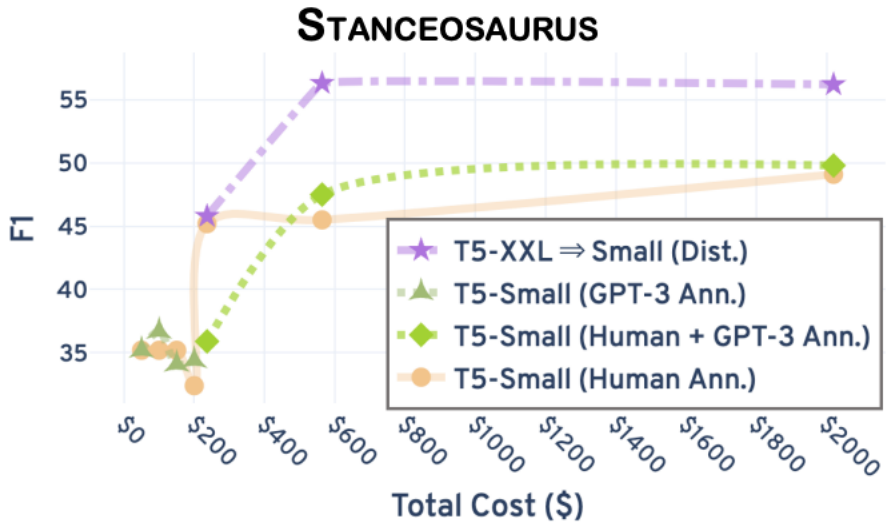
*Wang et al., 2022, Want To Reduce Labeling Cost? GPT-3 Can Help?

GPT-3.5 as an Annotator



*Wang et al., 2022, Want To Reduce Labeling Cost? GPT-3 Can Help?





Takeaways

Q. Given a limited budget,
how to invest it to train a compact model
in an economically efficient manner?

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

Takeaways

Q. Given a limited budget,
how to invest it to train a compact model
in an economically efficient manner?

✓ **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**

Dist.: $161 (81.0 F1) - max

Ann.: $1,980 (81.0 F1)

$17,443 (87.5 F1) - max

Takeaways

Q. Given a limited budget,
how to invest it to train a compact model
in an economically efficient manner?

✓ **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**

Takeaways

Q. Given a limited budget,
how to invest it to train a compact model
in an economically efficient manner?

✓ **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**

✓ **More details and analyses in the paper, such as different sizes of large & compact models**