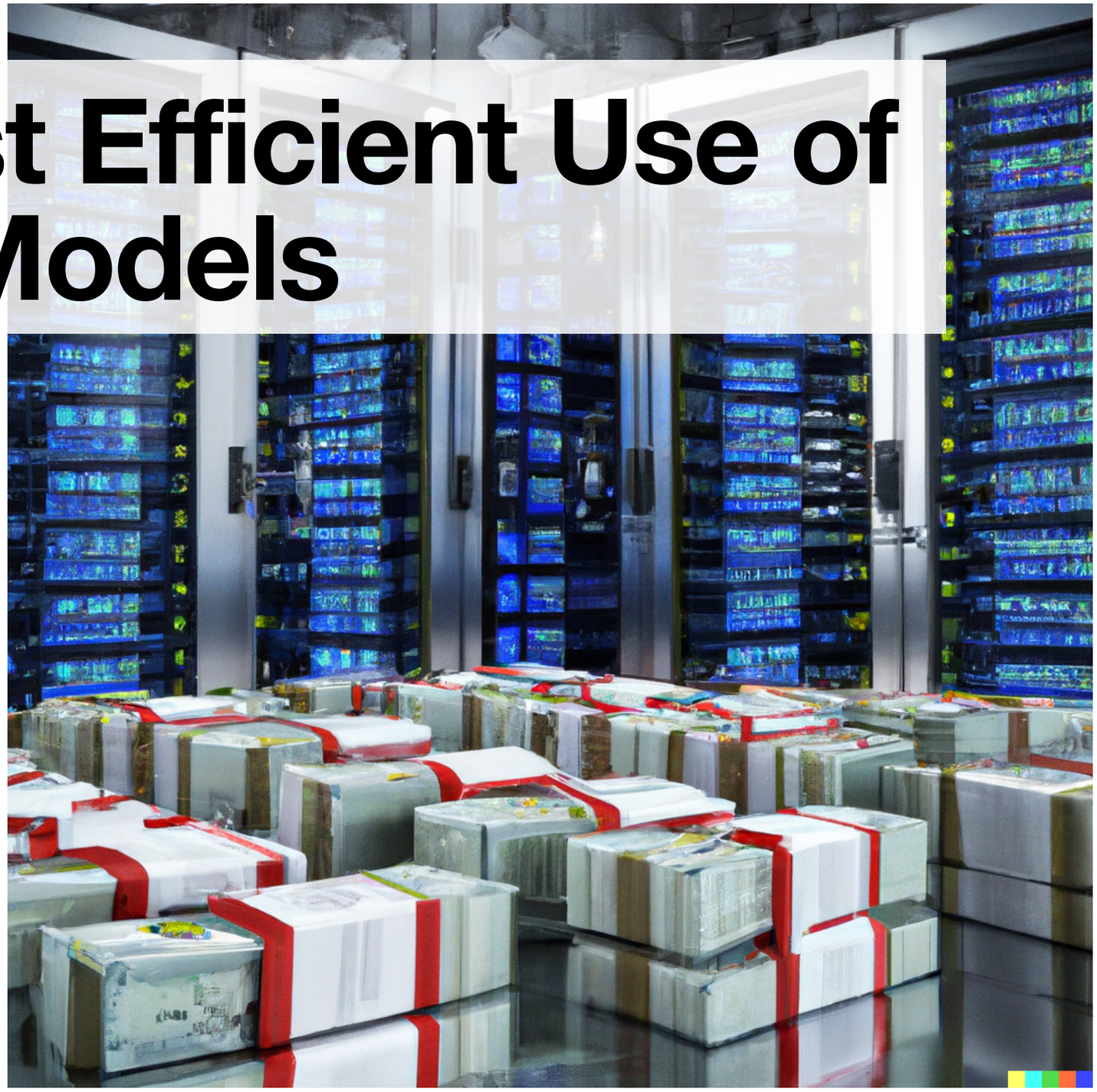
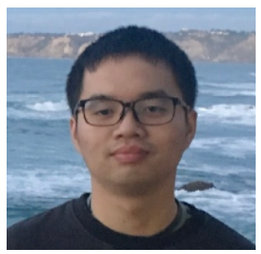
Georgia Tech Machine Learning

Towards Cost Efficient Use of Pre-trained Models

Alan Ritter



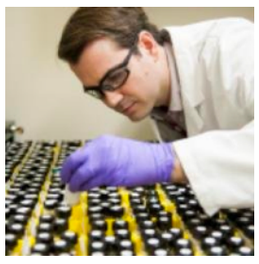
Joint Work With:



Fan Bai

Junmo Kang

Dayne Frietag



Peter Madrid

Gabriel Stanovsky

Wei Xu

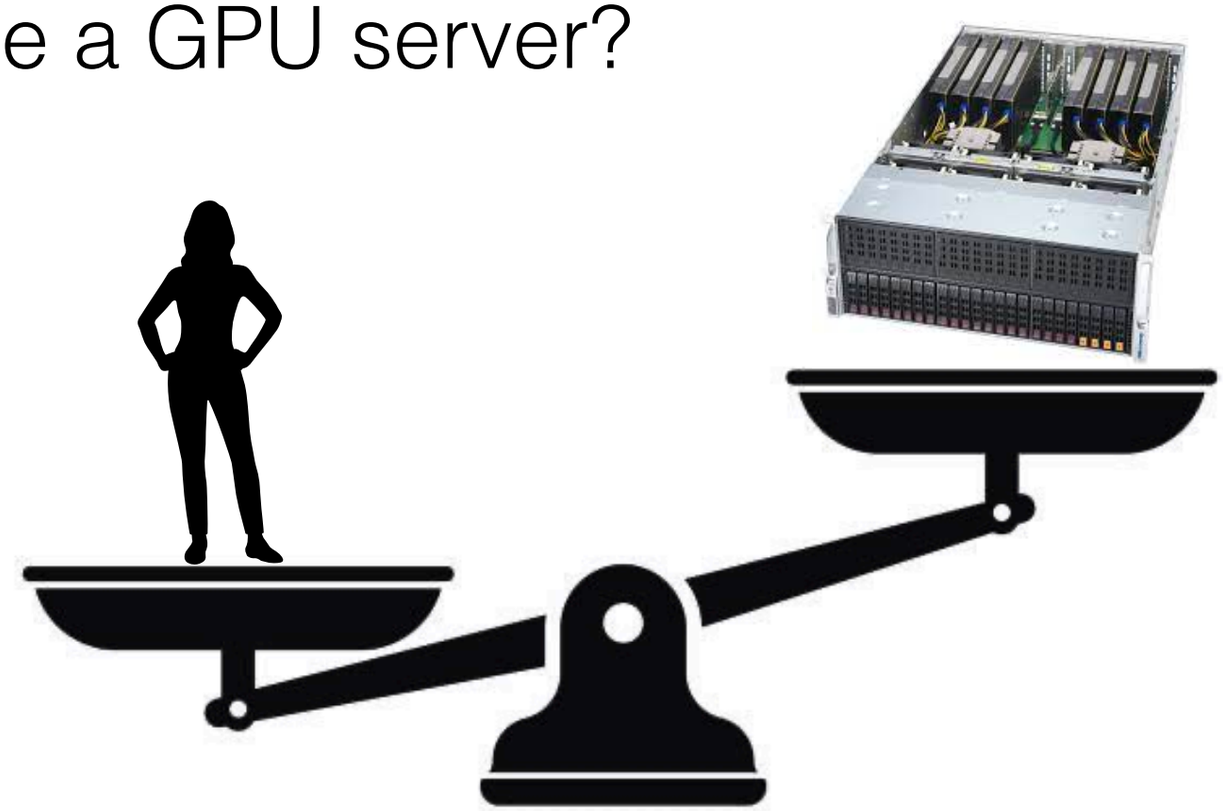
NLP is Getting Expensive!

- Reserved AWS **p4d.24xlarge** costs **$143,544/year!** 💰

NLP is Getting Expensive!

- Reserved AWS **p4d.24xlarge** costs **$143,544/year!** 💰

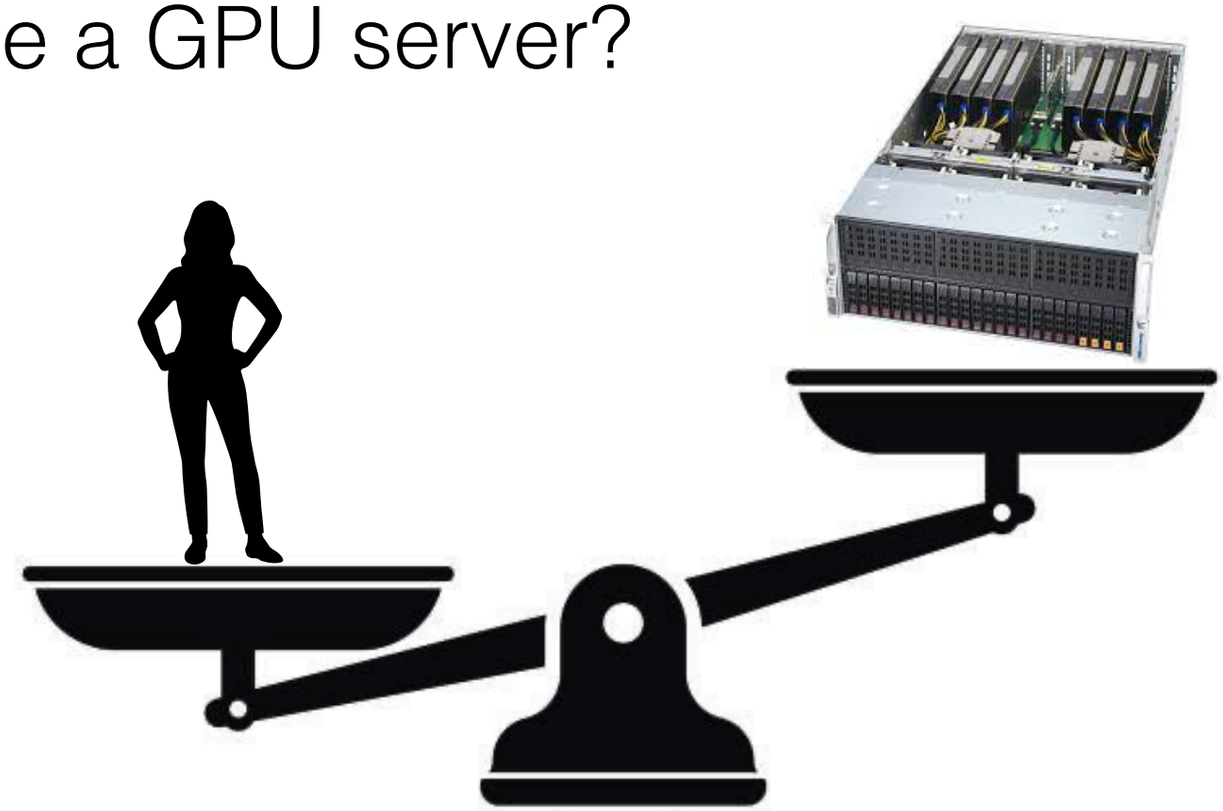
- Company: hire an employee?
  Or, purchase a GPU server?



NLP is Getting Expensive!

- Reserved AWS **p4d.24xlarge** costs **$143,544/year!** 💰
  - **Negative Externalities:**
    4993.2 kg Co2 ~1 car driving for a year (13K mi) 🚗💨

    Source: Machine Learning Emissions Calculator
- Company: hire an employee?
  Or, purchase a GPU server?



Energy and Policy Considerations for Deep Learning in NLP, Strubell et. al. 2019

NLP is Getting Expensive!

InstructGPT (Ouyang et. al. 2022)

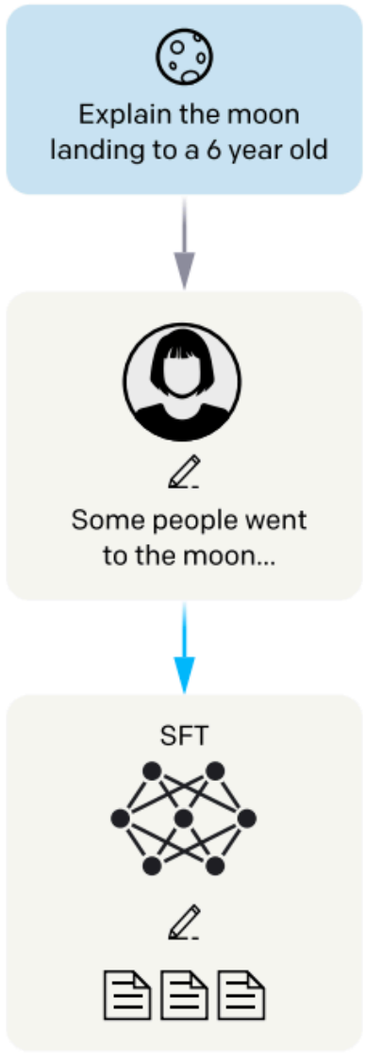
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



NLP is Getting Expensive!

InstructGPT (Ouyang et. al. 2022)

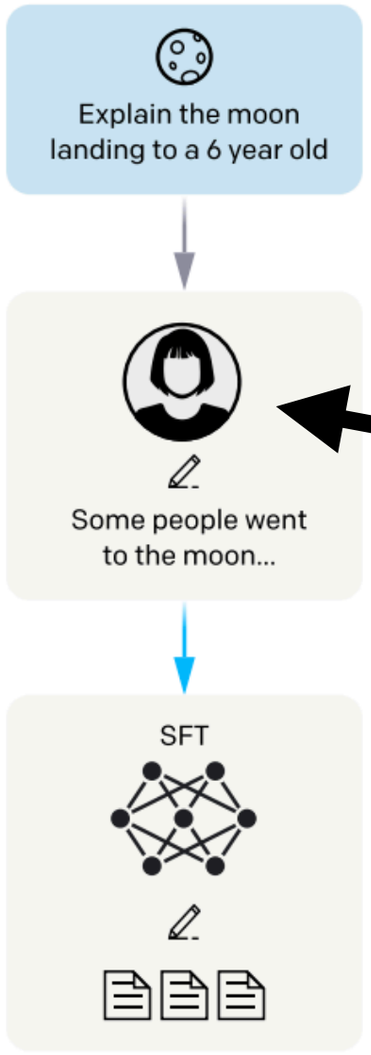
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



TECH

ChatGPT is powered by these contractors making $15 an hour

PUBLISHED SUN, MAY 7 2023•11:22 PM EDT

NBC NEWS David Ingram

NLP is Getting Expensive!

InstructGPT (Ouyang et. al. 2022)

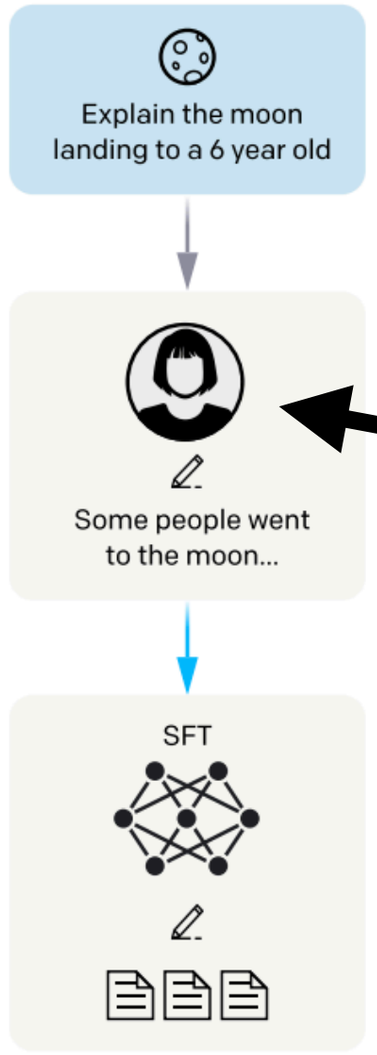
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



TECH

ChatGPT is powered by these contractors making $15 an hour

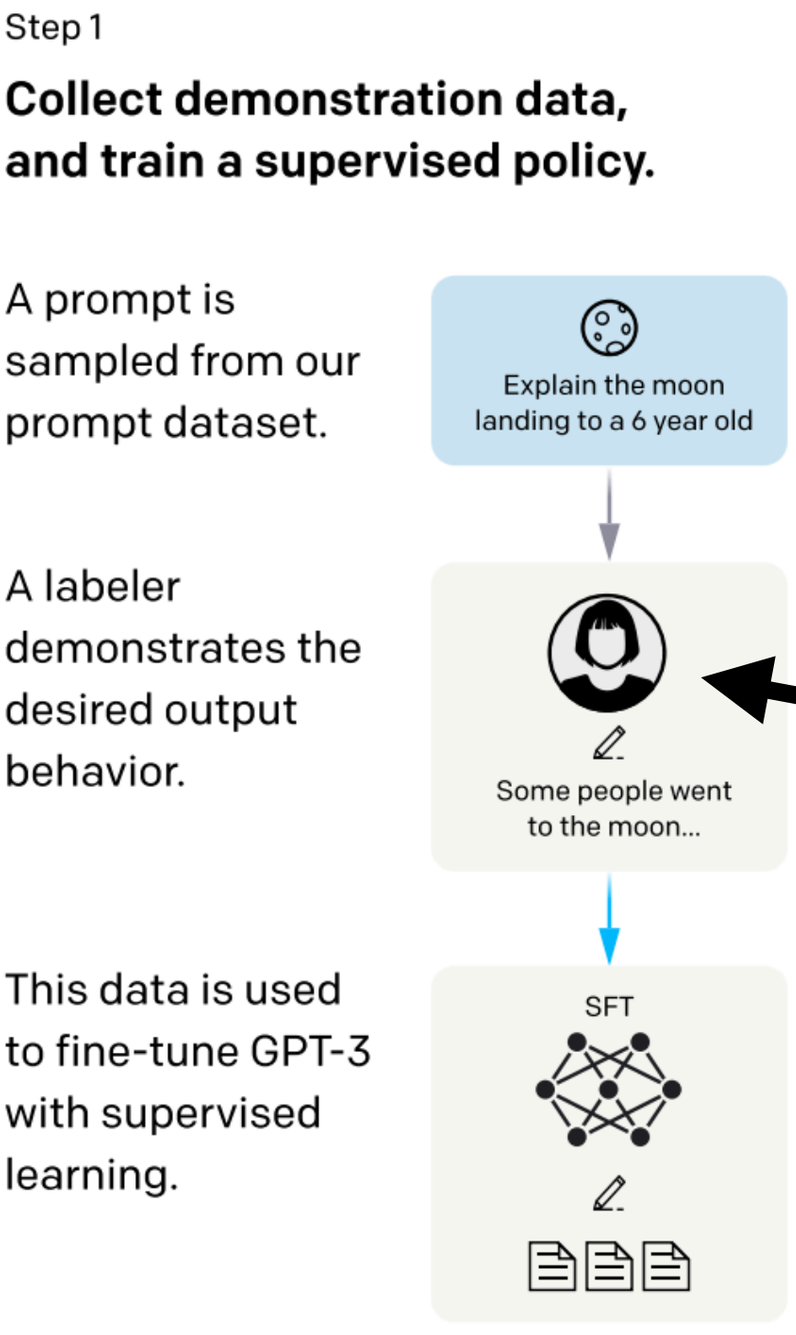
PUBLISHED SUN, MAY 7 2023•11:22 PM EDT

NBC NEWS David Ingram

GPT-3 was Estimated to cost $4 Million to Pre-Train

NLP is Getting Expensive!

InstructGPT (Ouyang et. al. 2022)



- **Computation vs. (Human) Annotation**

TECH

ChatGPT is powered by these contractors making $15 an hour

PUBLISHED SUN, MAY 7 2023•11:22 PM EDT

NBC NEWS David Ingram

GPT-3 was Estimated to cost $4 Million to Pre-Train

Tradeoff #1: **Domain Adaptive Pre-Training** (Gururangan et. al. 2020)

Pre-train or Annotate? Domain Adaptation with a Constrained Budget

Fan Bai, Alan Ritter, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{fan.bai, alan.ritter, wei.xu}@cc.gatech.edu

- **Computation vs. (Human) Annotation**

EMNLP 2021

▸Computation vs. (Human) Annotation

Tradeoff #1: **Domain Adaptive Pre-Training** (Gururangan et. al. 2020)

Pre-train or Annotate? Domain Adaptation with a Constrained Budget

Fan Bai, Alan Ritter, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{fan.bai, alan.ritter, wei.xu}@cc.gatech.edu

EMNLP 2021

▸Computation vs. (Human) Annotation

Tradeoff #1: **Domain Adaptive Pre-Training** (Gururangan et. al. 2020)

Pre-train or Annotate? Domain Adaptation with a Constrained Budget

Fan Bai, Alan Ritter, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{fan.bai, alan.ritter, wei.xu}@cc.gatech.edu

EMNLP 2021

Tradeoff #2: **Knowledge Distillation** (Sanh et. al. 2019, Zhou et. al. 2022)

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

Junmo Kang, Wei Xu, Alan Ritter
School of Interactive Computing
Georgia Institute of Technology
junmo.kang@gatech.edu {wei.xu, alan.ritter}@cc.gatech.edu

ACL 2023

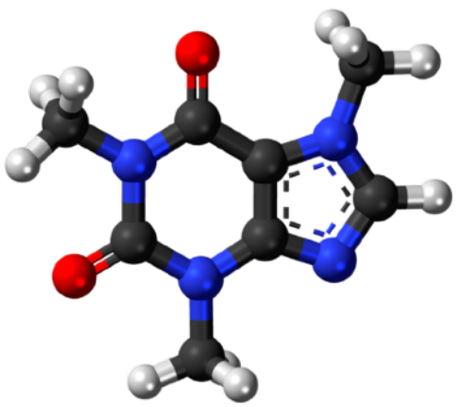
Computation vs. Annotation

Tradeoff #1: Pre-Train or Annotate

Domain #1

Existing NLP Model/Dataset

Domain #2



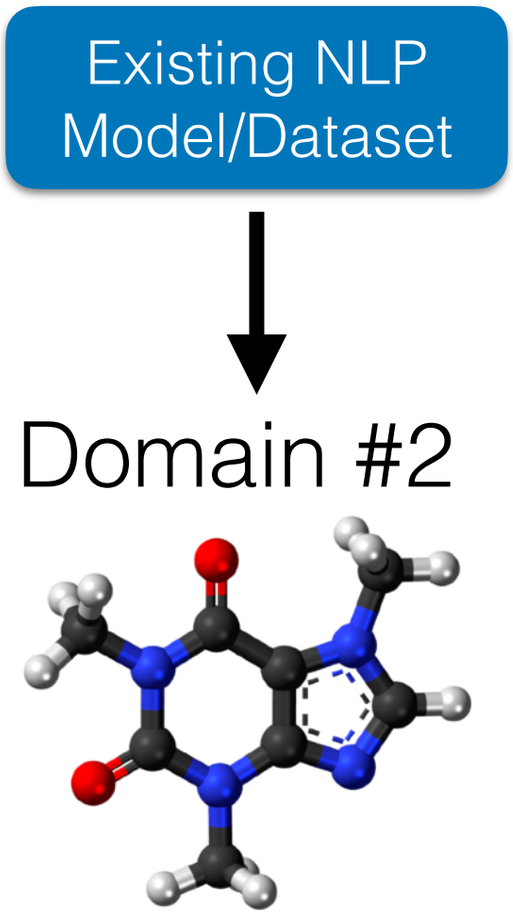
Computation vs. Annotation

Tradeoff #1: Pre-Train or Annotate

Domain Adaptive Pre-Training
(Han and Eisenstein, 2019)
(Gururangan et. al. 2020)

Domain #1

Existing NLP Model/Dataset



Domain #2

Hand-label data

Pre-train in-domain

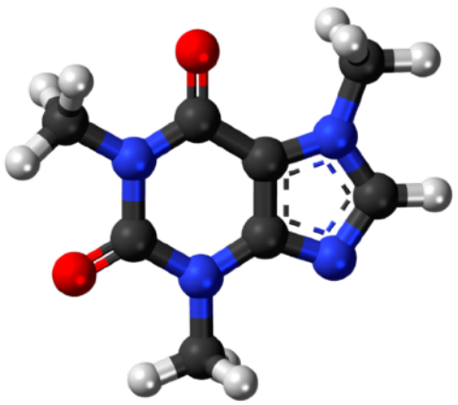
Computation vs. Annotation

Tradeoff #1: Pre-Train or Annotate

Domain #1

Existing NLP Model/Dataset

Domain #2



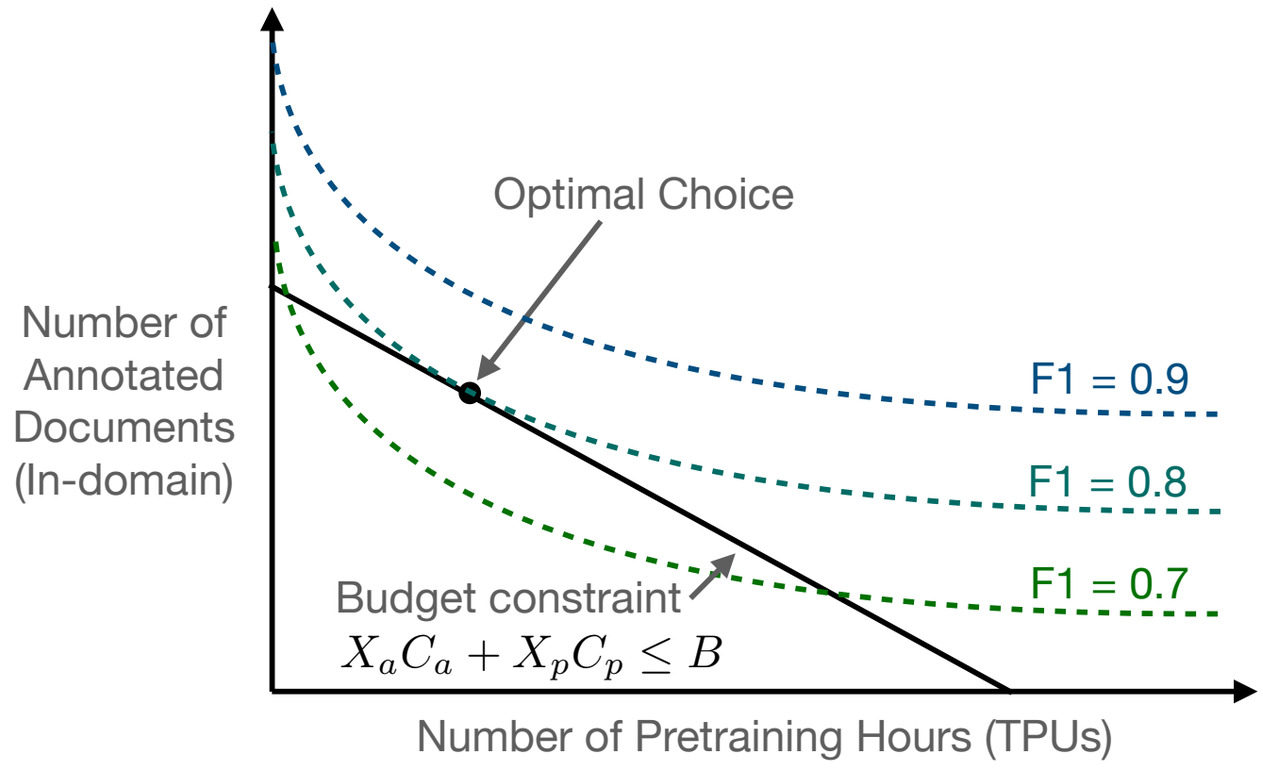
Domain Adaptive Pre-Training
(Han and Eisenstein, 2019)
(Gururangan et. al. 2020)

Hand-label data **Pre-train in-domain**

Conventional Wisdom: data labeling is expensive.

Q: Given current GPU/TPU costs, when is in-domain pre-training economical?

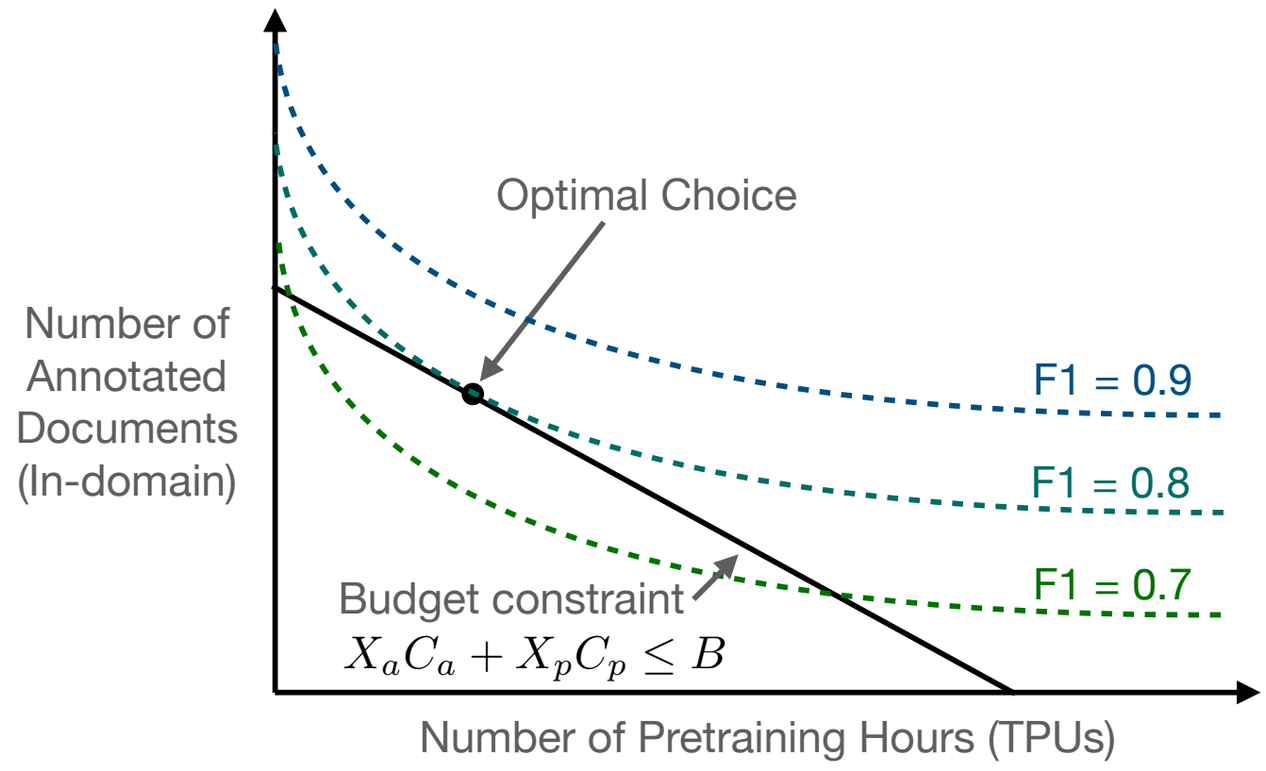
Key Insight: view domain adaptation through the lens of **consumer theory**



Q: Given current GPU/TPU costs, when is in-domain pre-training economical?

Q: Given current GPU/TPU costs, when is in-domain pre-training economical?

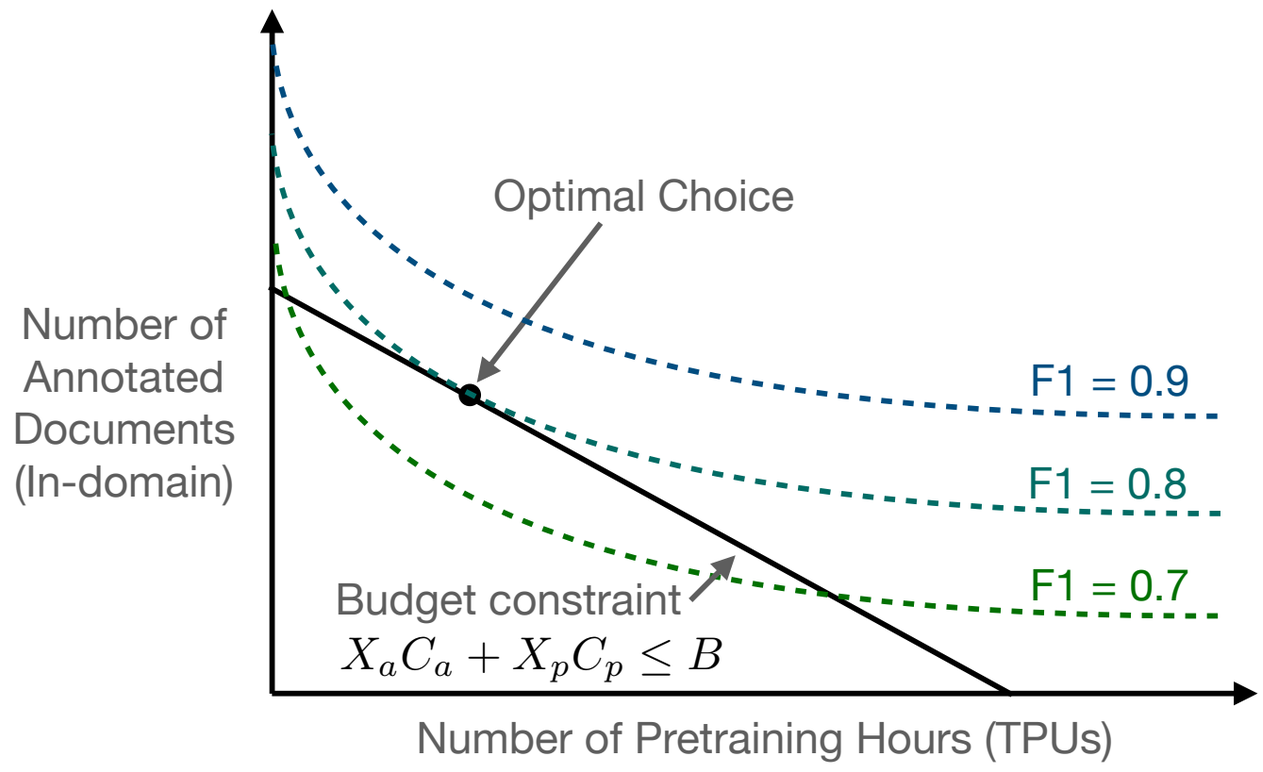
Key Insight: view domain adaptation through the lens of **consumer theory**



Q: Given current GPU/TPU costs, when is in-domain pre-training economical?

Key Insight: view domain adaptation through the lens of **consumer theory**

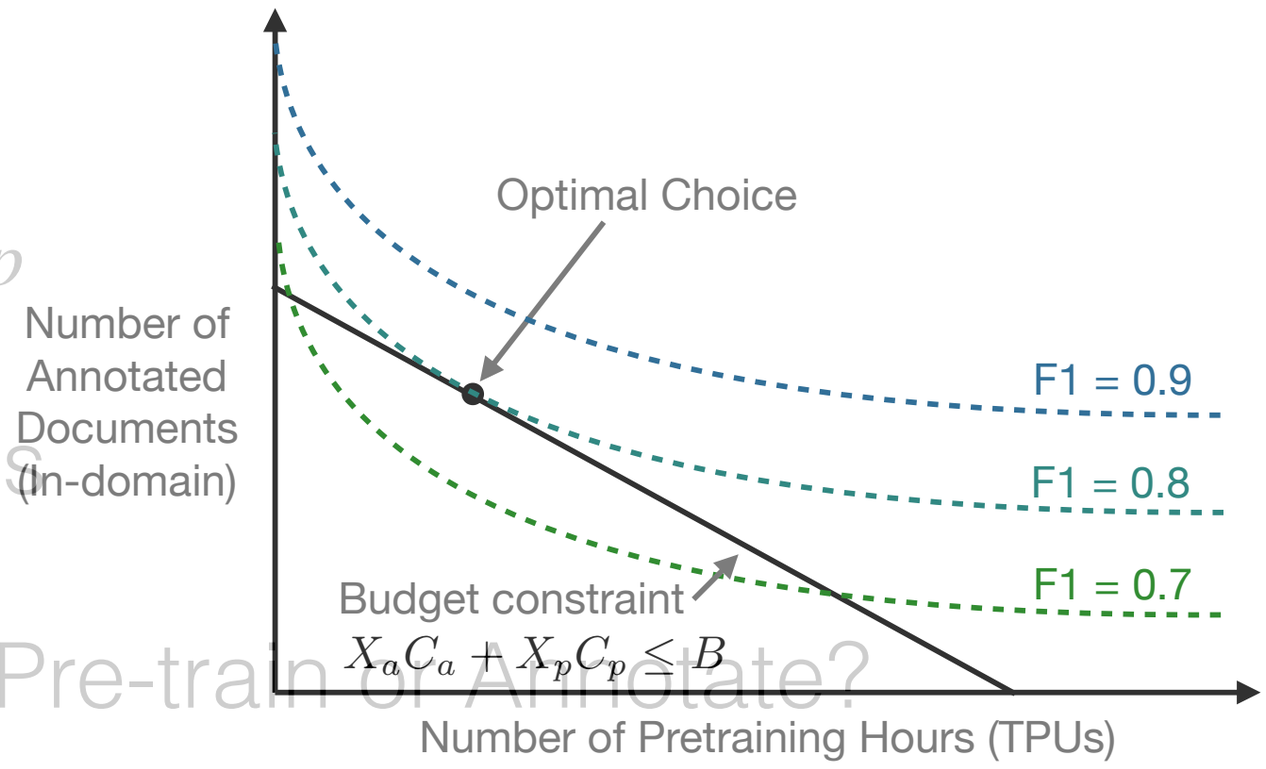
- ★ X_a annotated documents at a cost of C_a each
- ★ X_p hours of pre-training with cost C_p



Maximize utility, $U(X_a, X_p)$, within the budget constraint $X_a C_a + X_p C_p \leq B$.

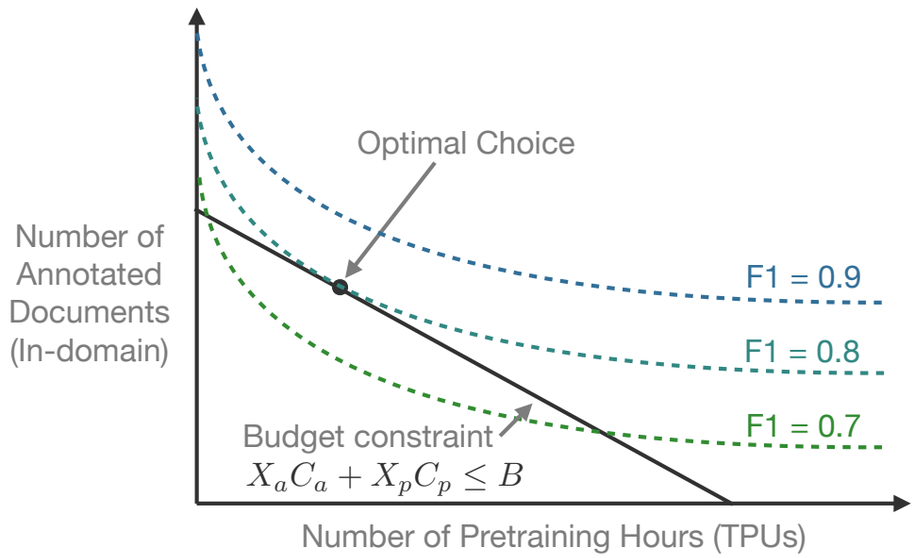
Remaining Questions

1. Annotation cost C_a
2. Pre-training cost C_p
3. Experimental details
4. Recommendation: Pre-train or Annotate?

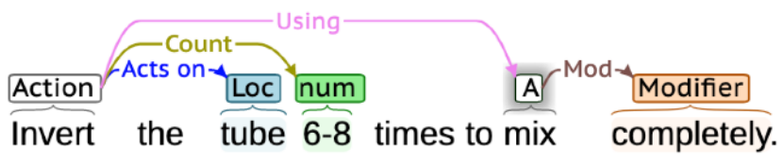
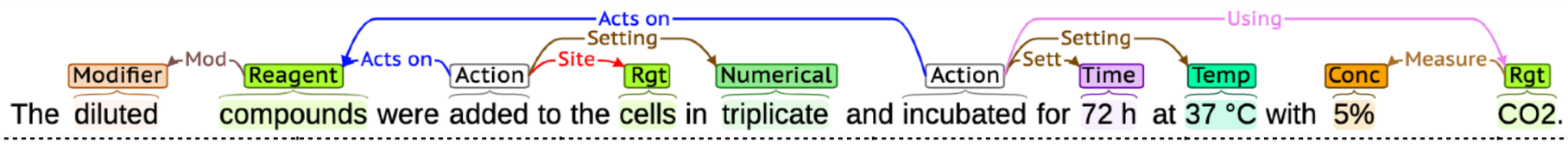
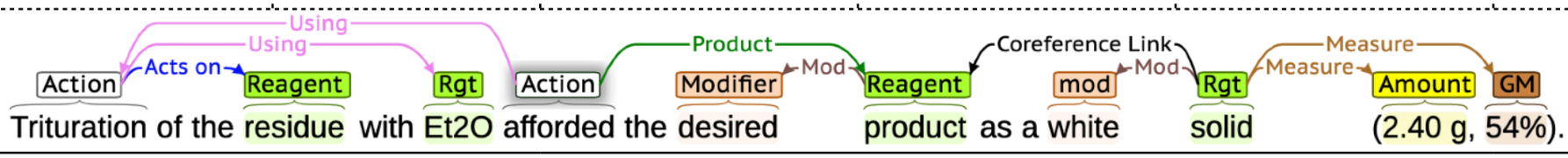


Remaining Questions

1. Annotation cost C_a
2. Pre-training cost C_p
3. Experimental details
4. Recommendation: Pre-train or Annotate?

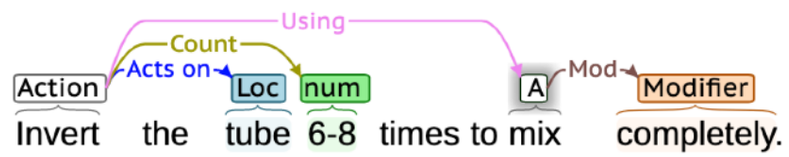
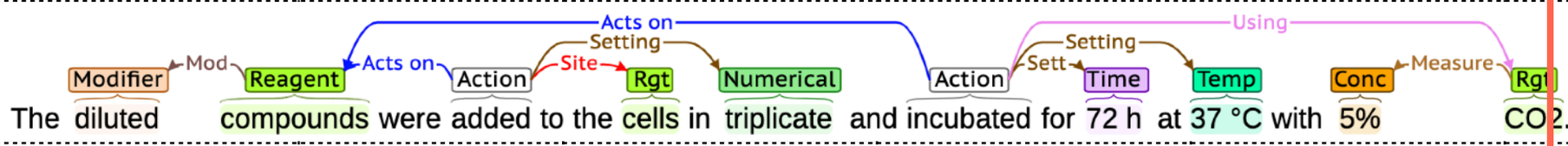
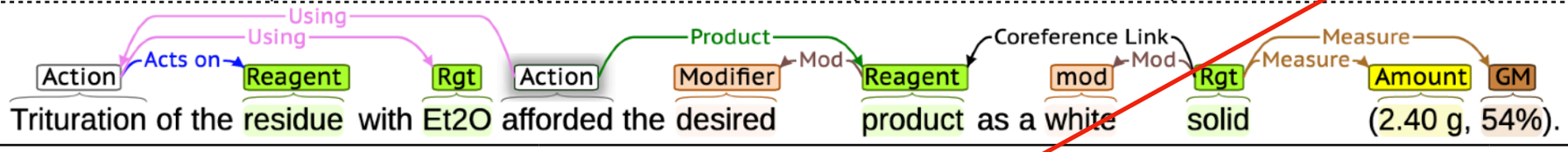


Estimating Annotation Cost, C_a

	Domain	#Procedures	#Sent.	Total Cost	Price/Sent.
Wet Lab Protocols Tabassum et al.	protocols	726	17,658	$7,820	**$0.44**
					
PubMed Materials & Methods (this work)	journal articles	191	1,699	$1,730	**$1.02**
					
Synthetic Procedures (this work)	chemistry	992	8,331	$5,000	**$0.60**
					

*Annotators were paid 13 USD/hour.

Estimating Annotation Cost, C_a

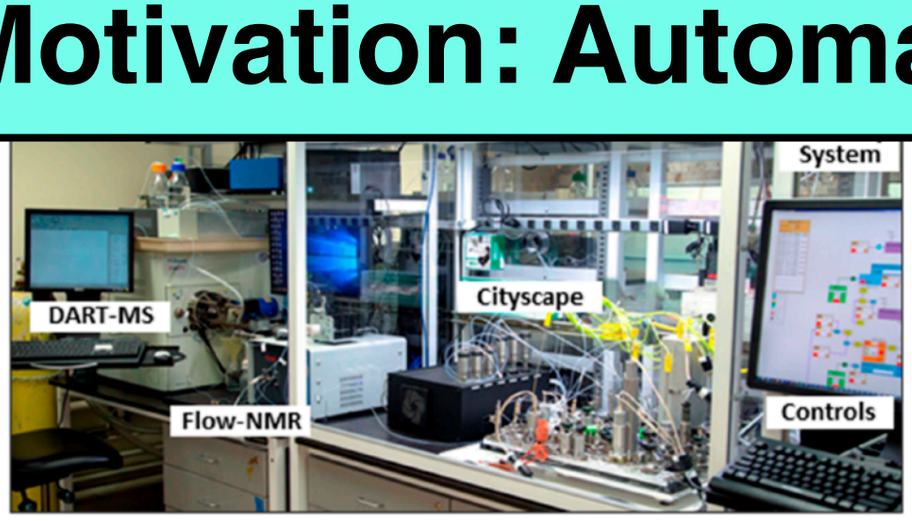
	Domain	#Procedures	#Sent.	Total Cost	Price/Sent.
Wet Lab Protocols Tabassum et al.	protocols	726	17,658	$7,820	**$0.44**
					
PubMed Materials & Methods (this work)	journal articles	191	1,699	$1,730	**$1.02**
					
Synthetic Procedures (this work)	chemistry	992	8,331	$5,000	**$0.60**
					

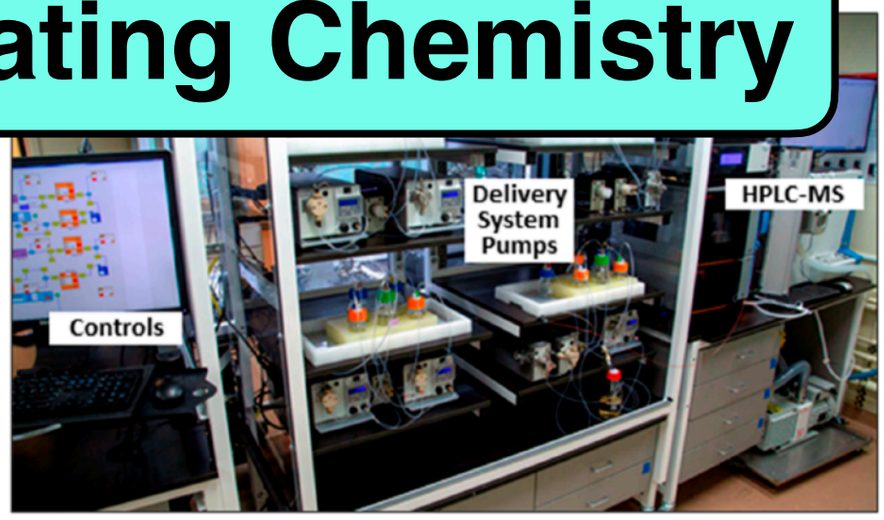
*Annotators were paid 13 USD/hour.

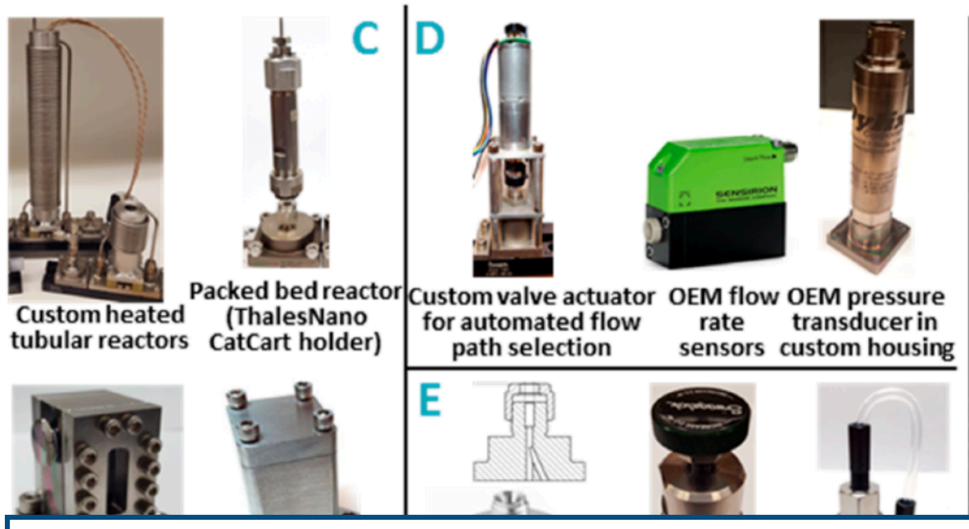
Annotation cost C_a (Price / Sent)

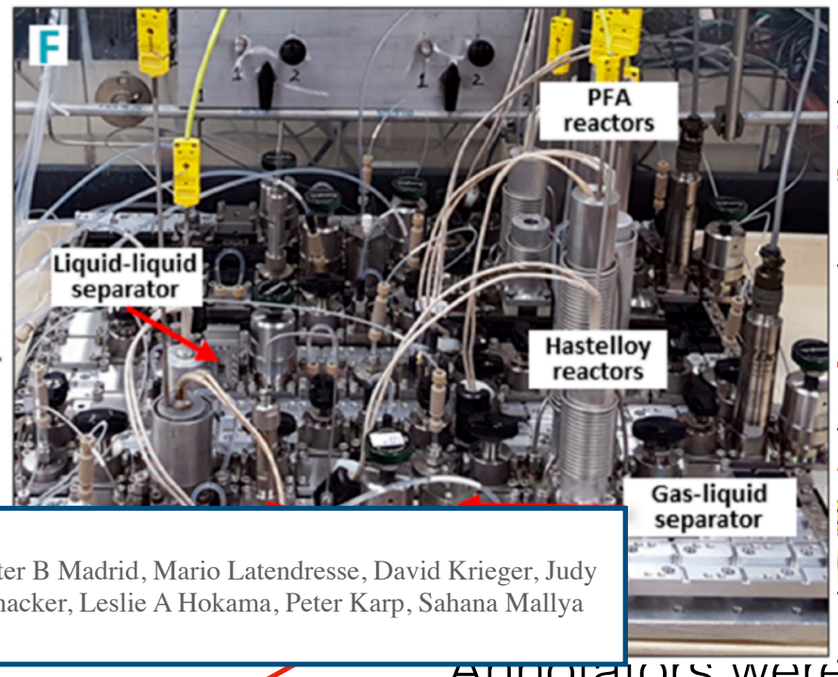
Estimating Annotation Cost, C_a

Motivation: Automating Chemistry









	Price/Sent.
	$0.44
	$1.02
	$0.60

easure Rgt CO2.

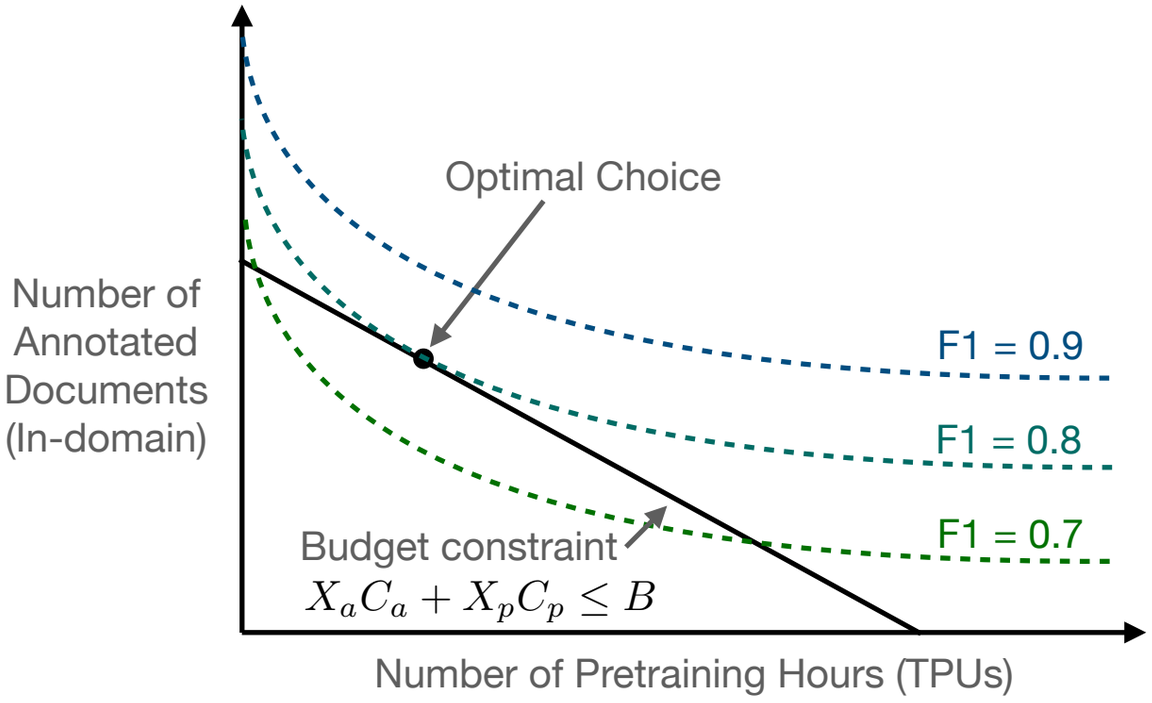
ount GM 0 g, 54%).

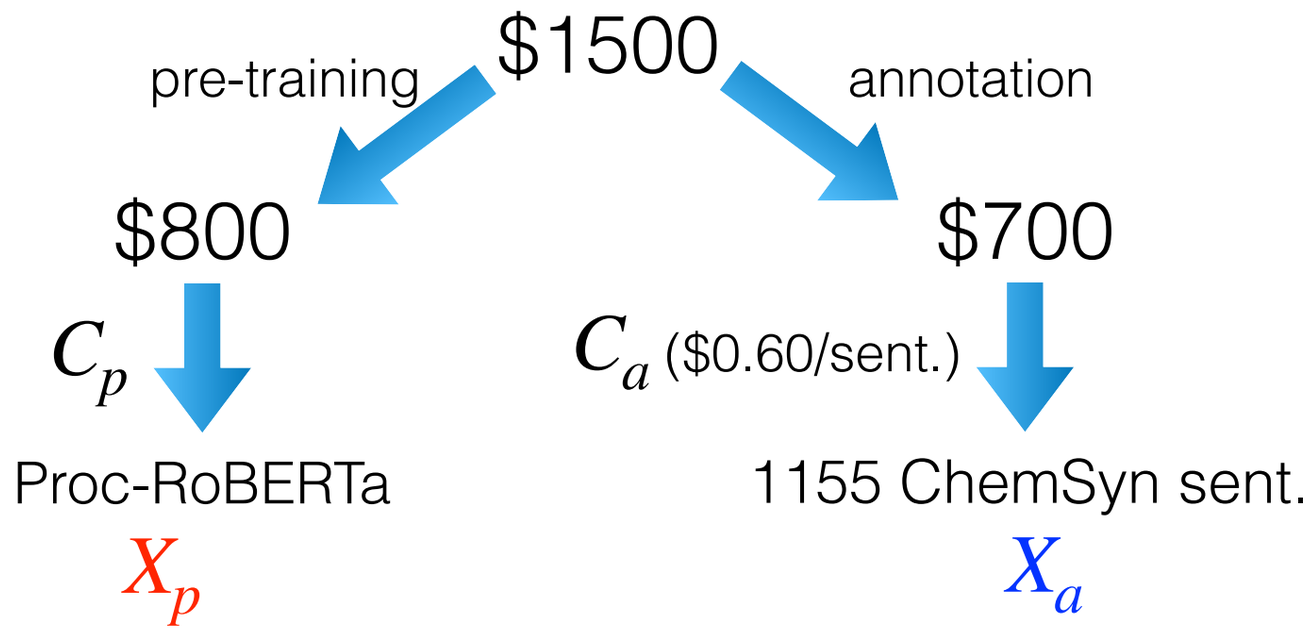
Fully automated chemical synthesis: toward the universal synthesizer
Nathan Collins, David Stout, Jin-Ping Lim, Jeremiah P Malerich, Jason D White, Peter B Madrid, Mario Latendresse, David Krieger, Judy Szeto, Vi-Anh Vu, Kristina Rucker, Michael Deleo, Yonael Gorfu, Markus Krummenacker, Leslie A Hokama, Peter Karp, Sahana Mallya
Organic Process Research & Development

Annotators were paid 13 USD/hour.

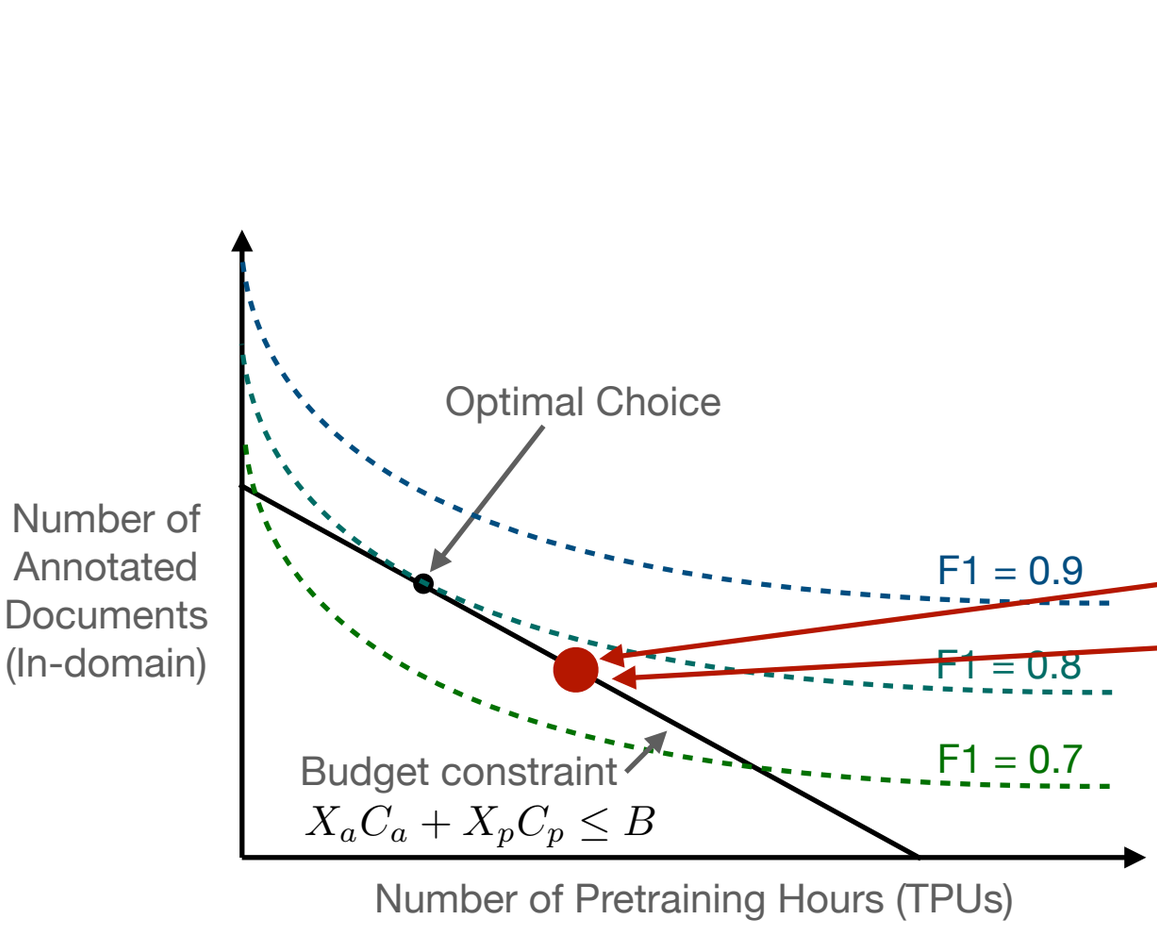
Annotation cost C_a (Price / Sent)

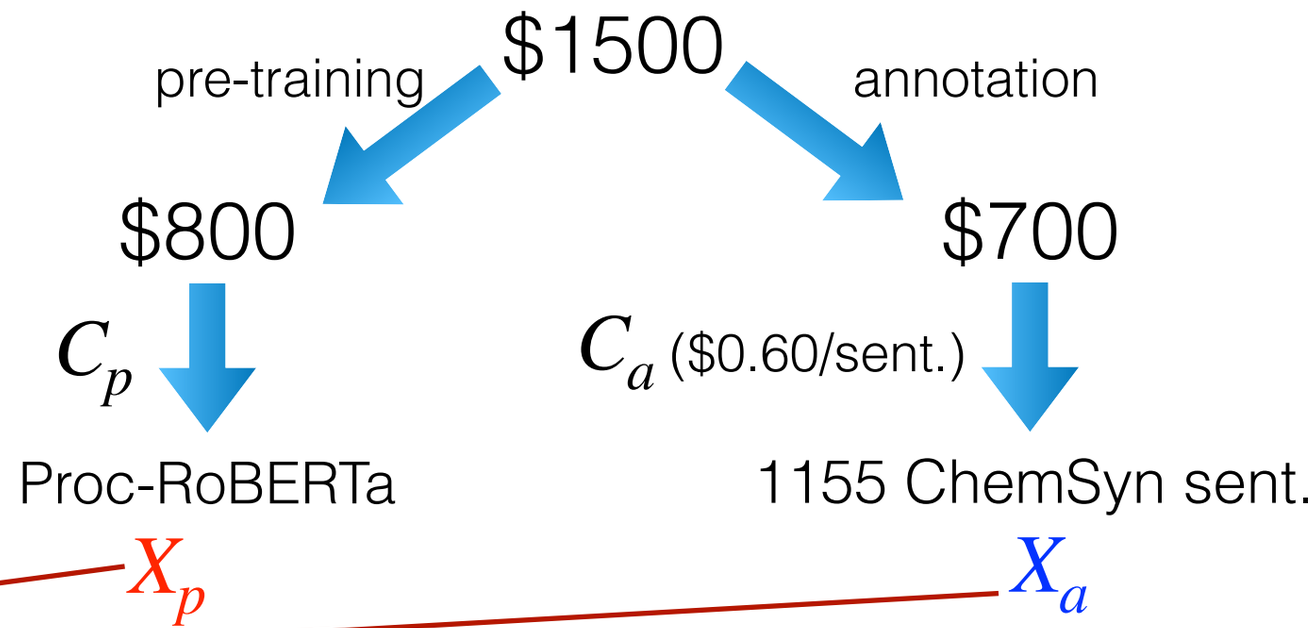
Measuring Utility, $U(X_a, X_p)$



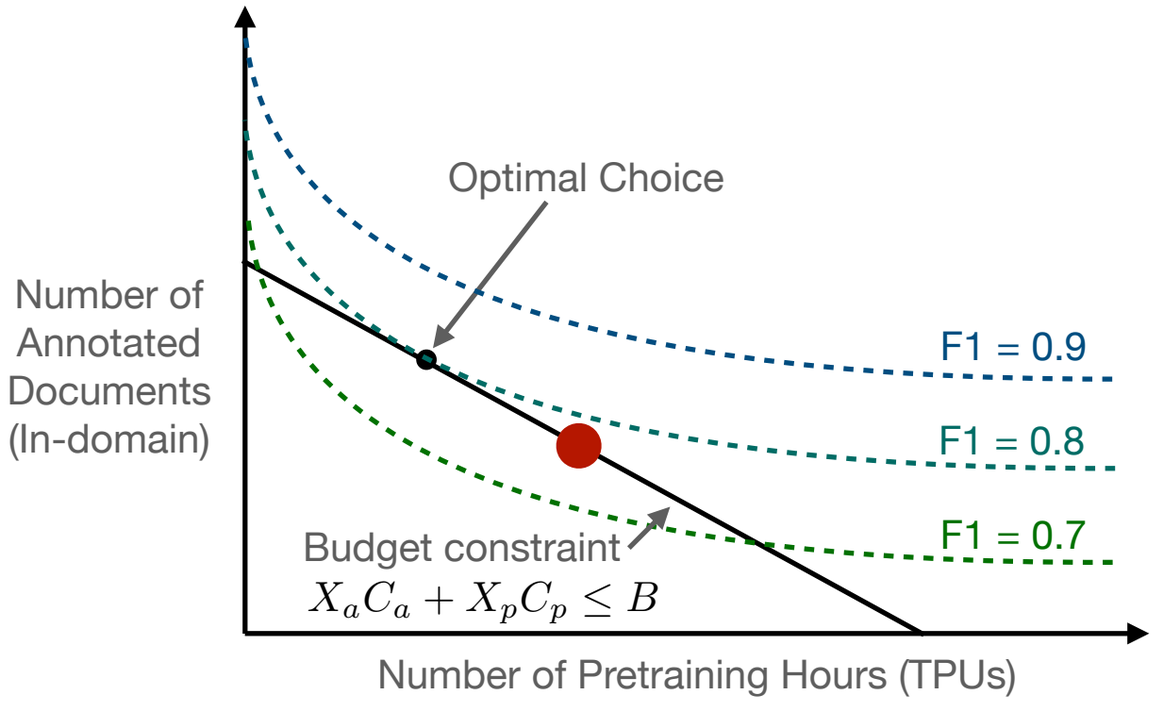


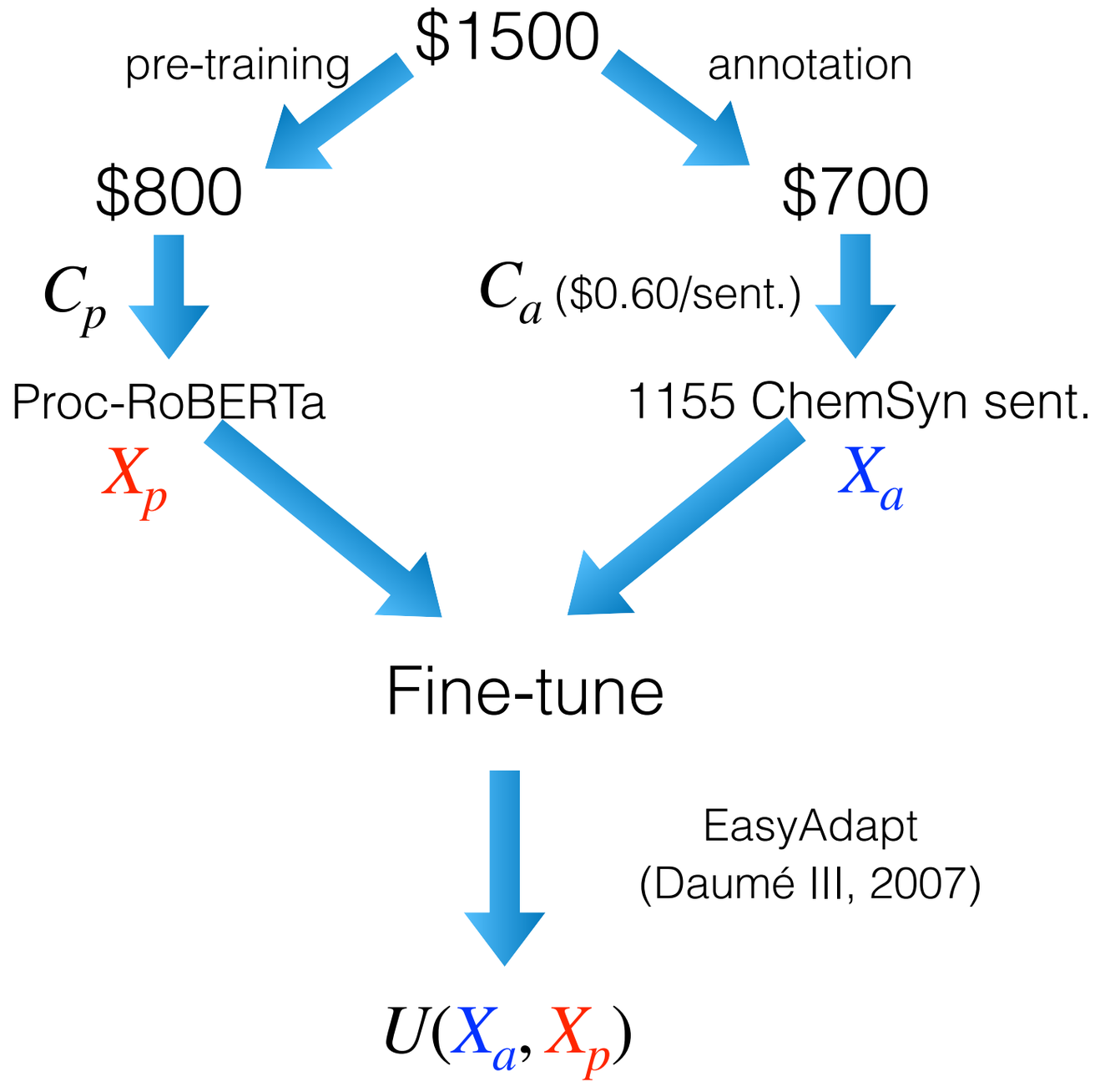
Measuring Utility, $U(X_a, X_p)$





Measuring Utility, $U(X_a, X_p)$

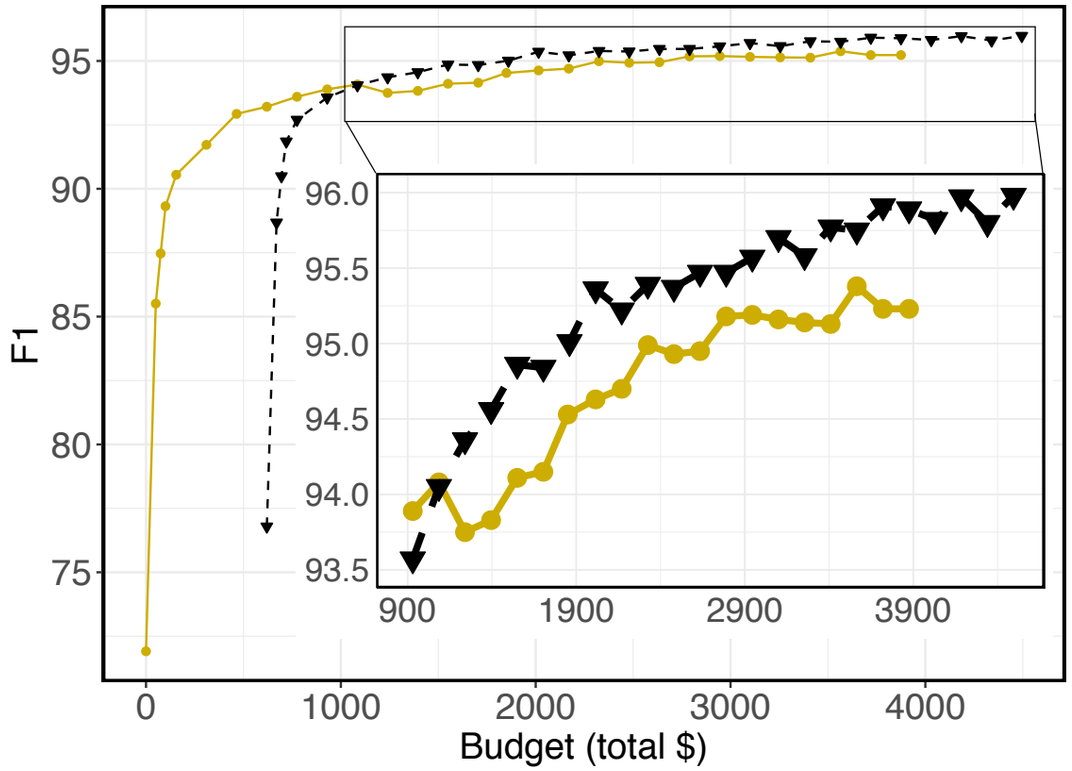




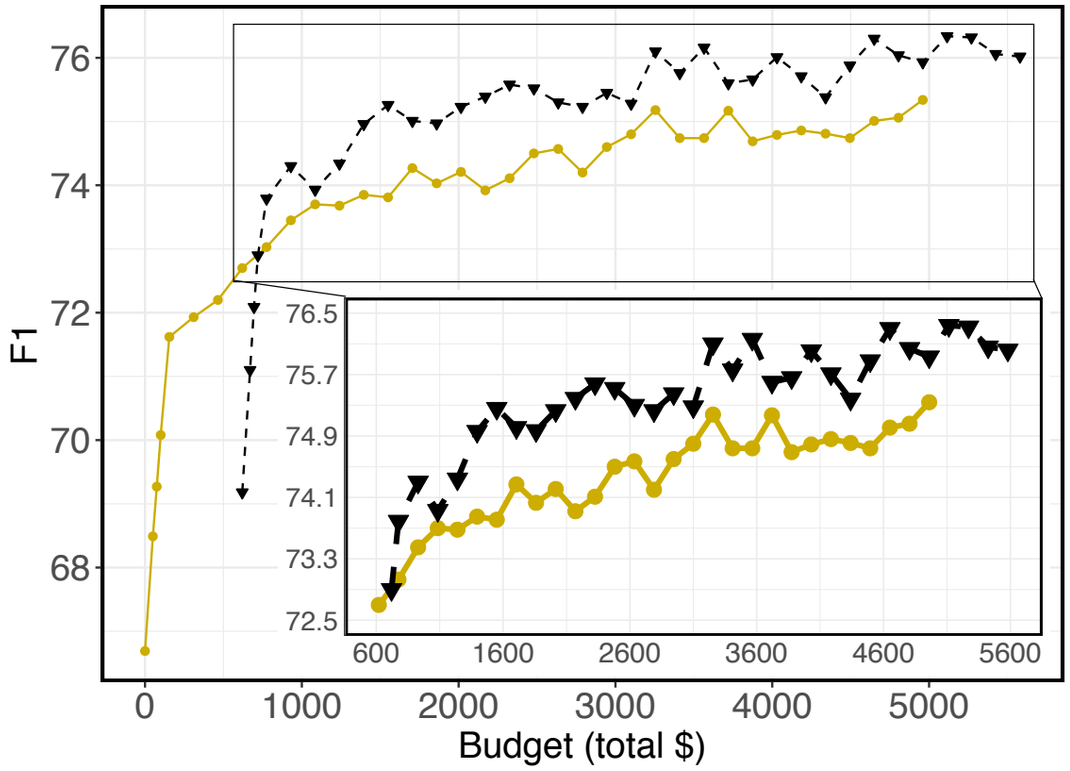
Pre-train or Annotate?

: Allocate full budget to annotation

: Pre-train in-domain model, spend remaining $ on annotation



PubMed ⇒ Synthetic Procedures

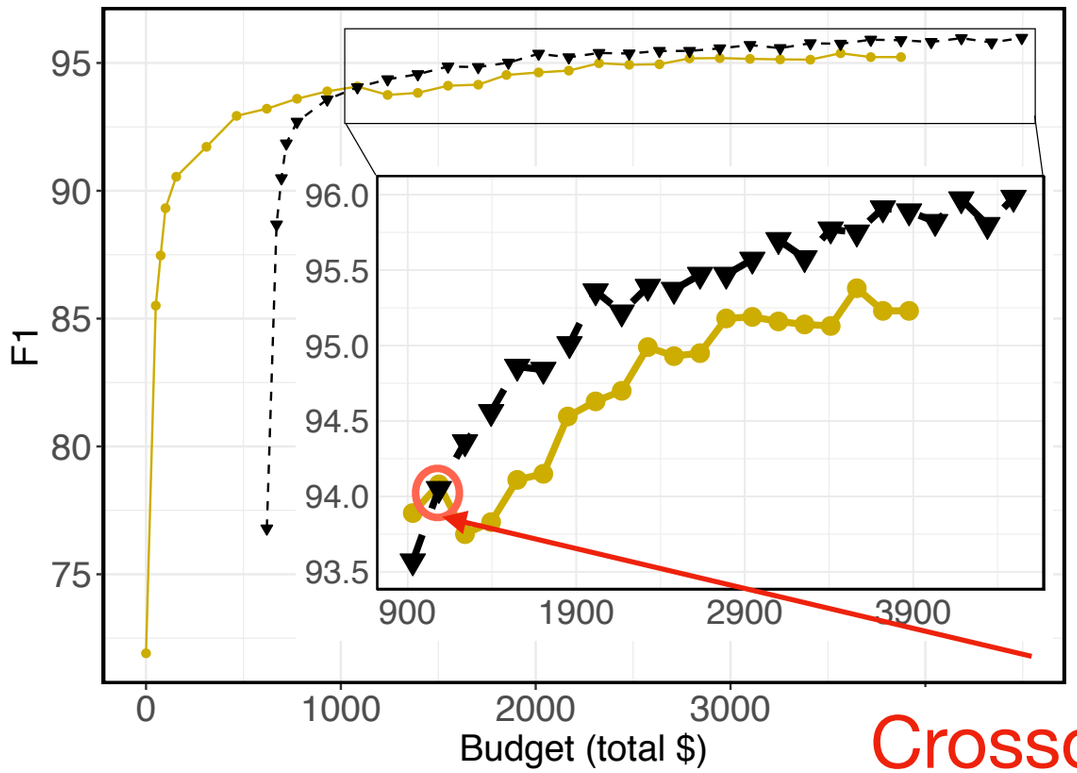


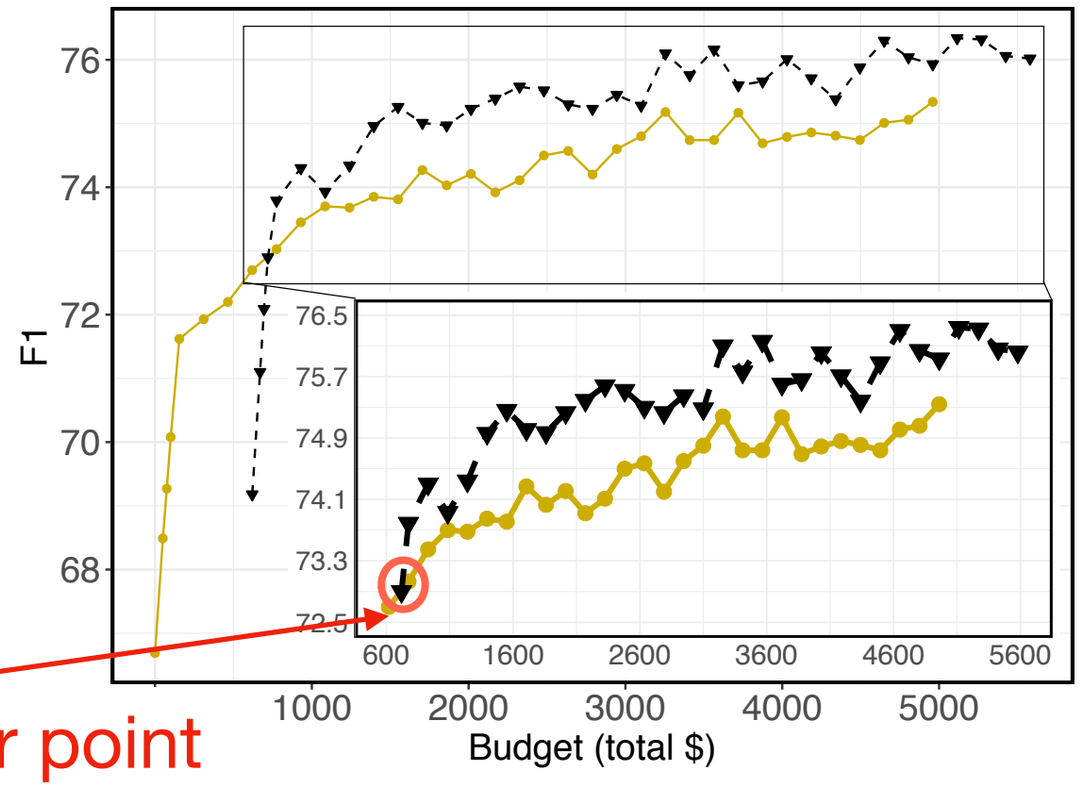
Synthetic Procedures ⇒ WLP

Pre-train or Annotate?

: Allocate full budget to annotation

: Pre-train in-domain model, spend remaining $ on annotation





Crossover point

PubMed

Recommendation:

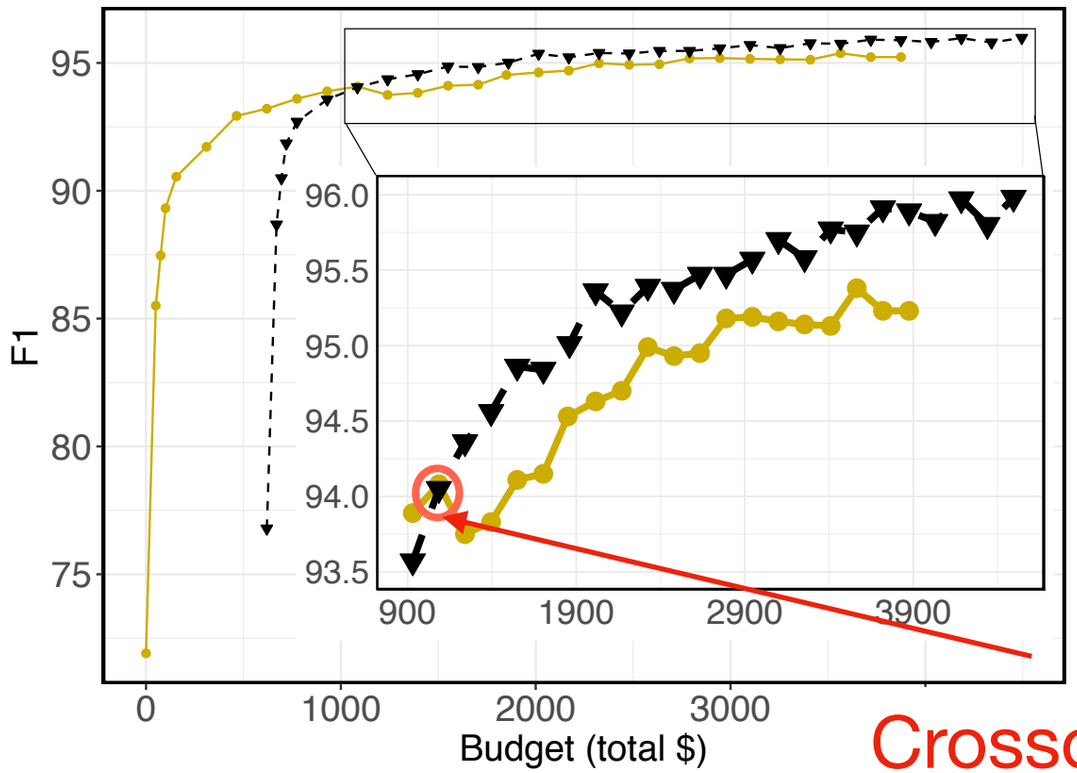
- Small Budget => Annotation
- Large Budget => Annotation + Pre-training

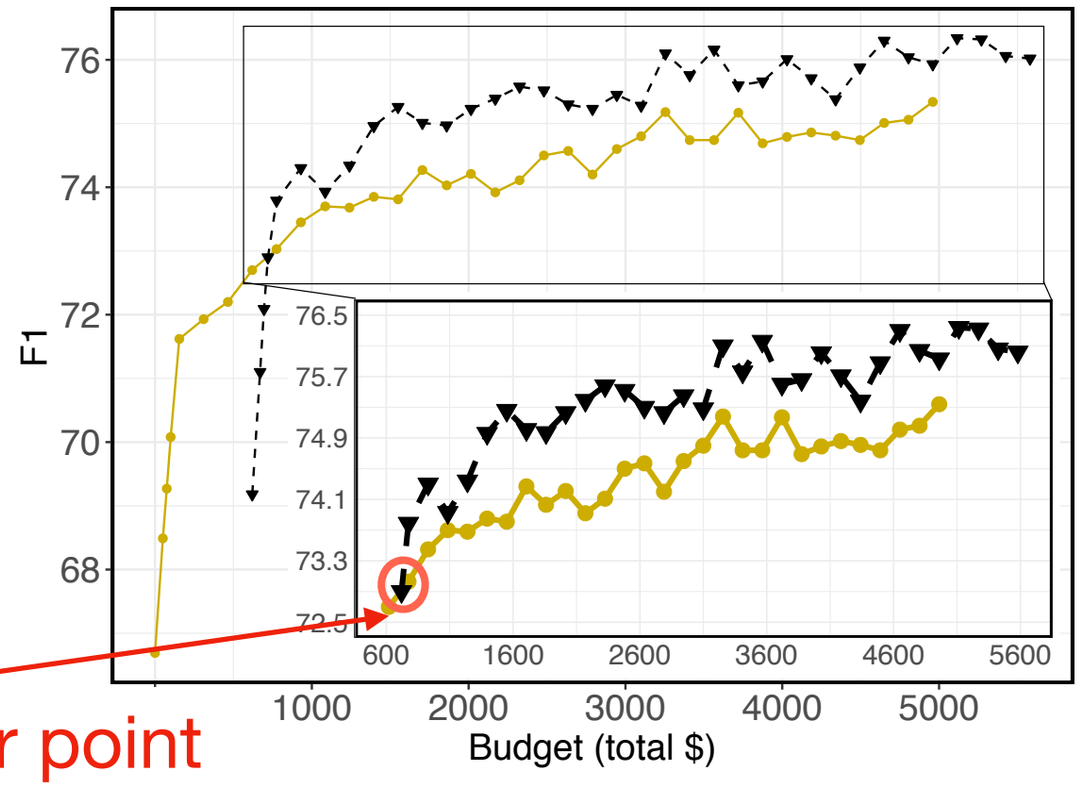
es ⇒ WLP

Pre-train or Annotate?

: Allocate full budget to annotation

: Pre-train in-domain model, spend remaining $ on annotation





Crossover point

PubMed

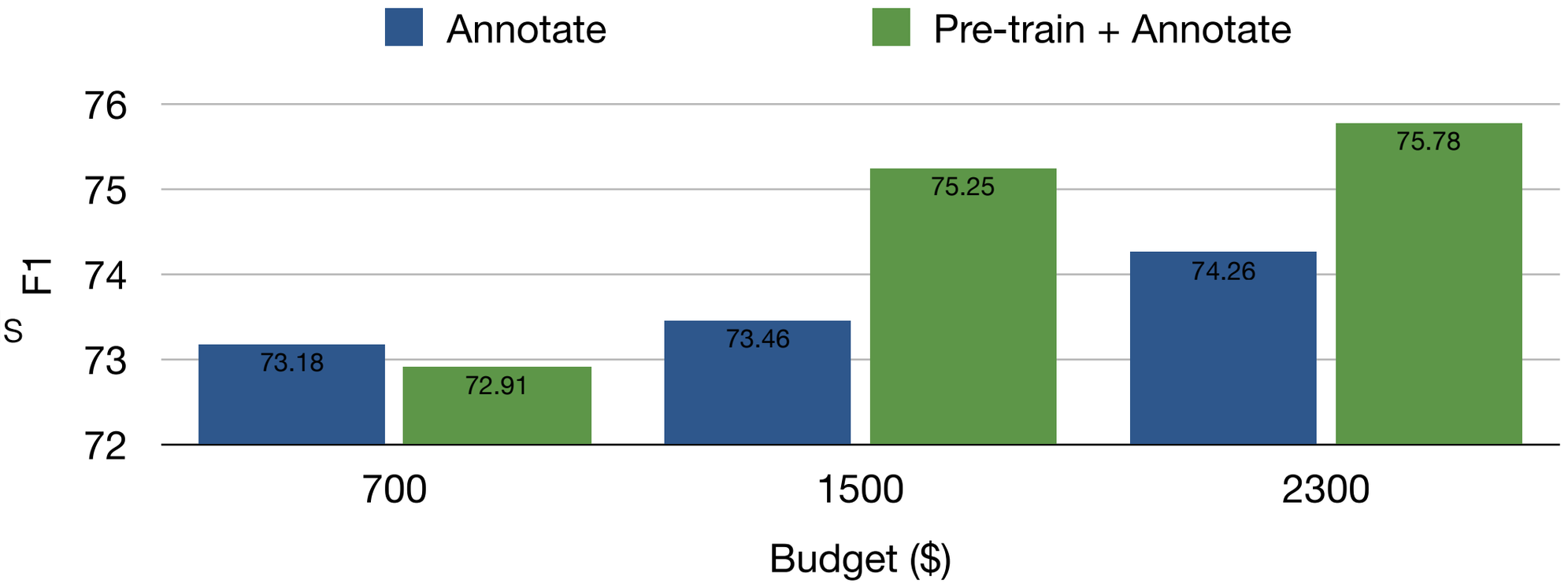
es⇒ WLP

Recommendation:

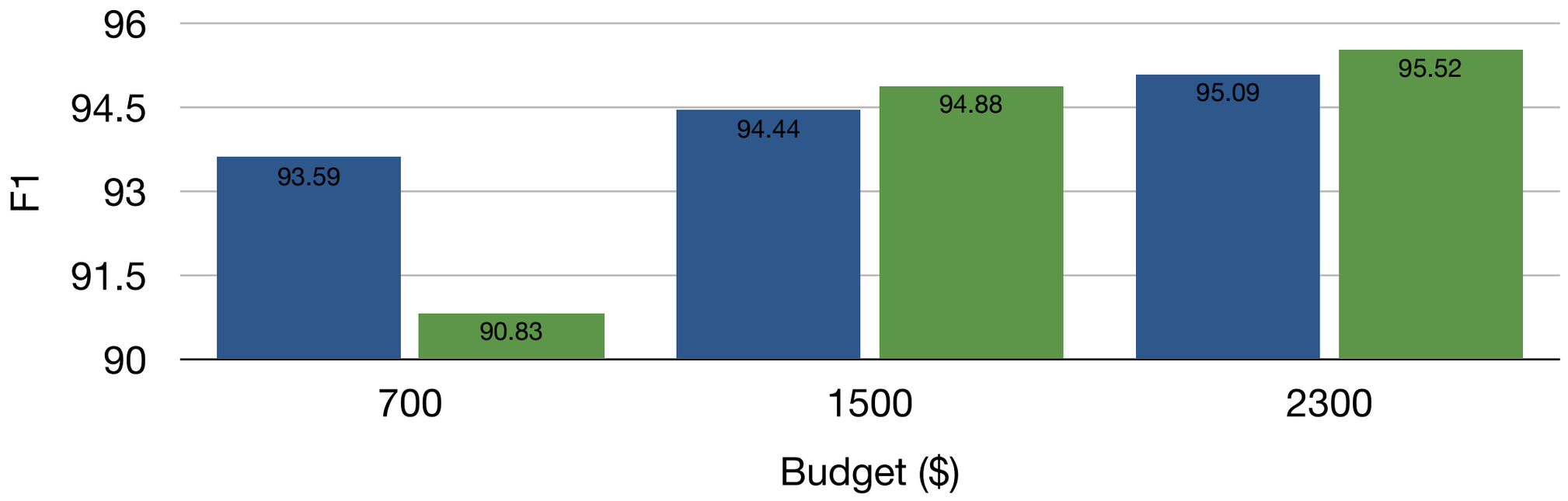
- Small Budget => Annotation
- Large Budget => Annotation + Pre-training
- **Annotation is more important, if you must choose**

Results are Similar Across Multiple Domains

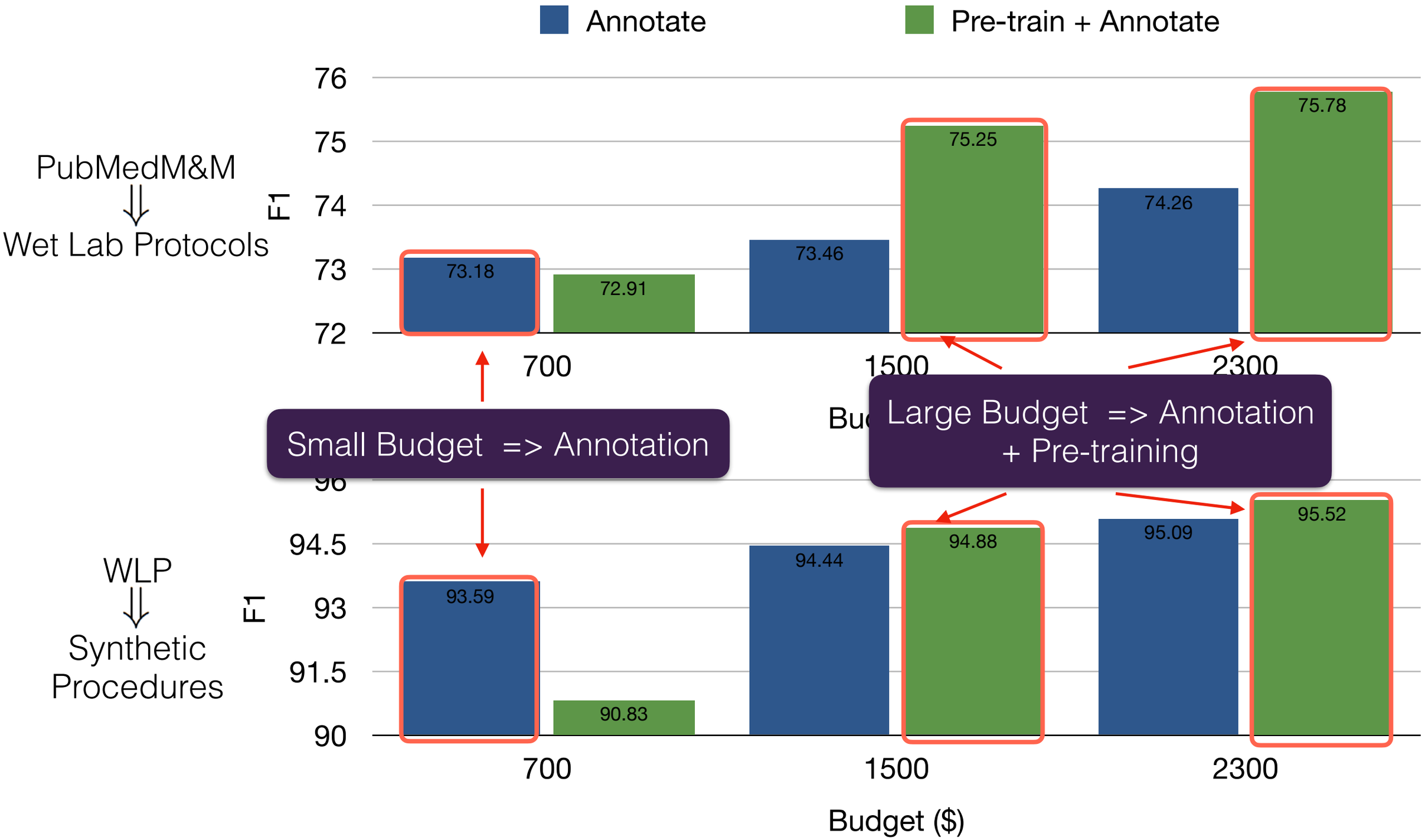
PubMedM&M
⇓
Wet Lab Protocols



WLP
⇓
Synthetic
Procedures



Results are Similar Across Multiple Domains



Demo: Synthetic Protocol Search

Synthesis Procedures

UNITED STATES PATENT AND TRADEMARK OFFICE uspto + European Patent Office **(6 million synth. procedures)**

US06169084 - Preparation of (2-methyl-4-(4-methyl-1-piperazinyl)-10H-thieno[2,3-b][1,5]benzodiazepine)dihydrate E.

1. A 0.5 g sample of technical grade olanzapine was suspended in ethyl acetate (10 mL) and toluene (0.6 mL).
2. The mixture was heated to 80° C. until all the solids dissolved.
3. The solution was cooled to 60° C. and water (1 mL) was added slowly.

...

7. TGA mass loss was 10.5%.
8. Yield: 0.3 g.

Demo: Synthetic Protocol Search

Synthesis Procedures

UNITED STATES PATENT AND TRADEMARK OFFICE uspto + European Patent Office **(6 million synth. procedures)**

US06169084 - Preparation of (2-methyl-4-(4-methyl-1-piperazinyl)-10H-thieno[2,3-b][1,5]benzodiazepine)dihydrate E.

1. A 0.5 g sample of technical grade olanzapine was suspended in ethyl acetate (10 mL) and toluene (0.6 mL).
2. The mixture was heated to 80° C. until all the solids dissolved.
3. The solution was cooled to 60° C. and water (1 mL) was added slowly.

   ...
7. TGA mass loss was 10.5%.
8. Yield: 0.3 g.

Extracted Information

1. Shallow Semantic Parsing (operation-level)

2. Slot Filling (procedure-level)

Patent ID	Reagent	Solvent	Product	Yield	...
US06169084	olanzapine	ethyl acetate / toluene	(2-methyl-4-(4-methyl..)	0.3 g	...
...	...	...	...	...	....

Demo: Synthetic Protocol Search

Synthesis Procedures

UNITED STATES PATENT AND TRADEMARK OFFICE uspto + European Patent Office **(6 million synth. procedures)**

US06169084 - Preparation of (2-methyl-4-(4-methyl-1-piperazinyl)-10H-thieno[2,3-b][1,5]benzodiazepine)dihydrate E.

1. A 0.5 g sample of technical grade olanzapine was suspended in ethyl acetate (10 mL) and toluene (0.6 mL).
2. The mixture was heated to 80° C. until all the solids dissolved.
3. The solution was cooled to 60° C. and water (1 mL) was added slowly.

…

7. TGA mass loss was 10.5%.
8. Yield: 0.3 g.

Semantic Search

Q: What **solvents** are used in the reactions that produce **(2-methyl-4-(4-methyl-1-piperazinyl**…?

Patent ID	Solvent	Full Procedure
US06169084	ethyl acetate / toluene	A 0.5 g sample of…
…	…	…

SYNKB

Q: What **amount** is **ethyl acetate** at when used to suspend **olanzapine**?

Patent ID	Amount	Full Procedure
US06169084	10 mL	A 0.5 g sample of…
…	…	…

SYNKB

Extracted Information

1. Shallow Semantic Parsing (operation-level)

2. Slot Filling (procedure-level)

Patent ID	Reagent	Solvent	Product	Yield	…
US06169084	olanzapine	ethyl acetate / toluene	(2-methyl-4-(4-methyl..)	0.3 g	…
…	…	…	…	…	….

Demo: Synthetic Protocol Search

Synthesis Procedures

UNITED STATES PATENT AND TRADEMARK OFFICE uspto + European Patent Office **(6 million synth. procedures)**

US06169084 - Preparation of (2-methyl-4-(4-methyl-1-piperazinyl)-10H-thieno[2,3-b][1,5]benzodiazepine)dihydrate E.

1. A 0.5 g sample of technical grade olanzapine was suspended in ethyl acetate (10 mL) and toluene (0.6 mL).
2. The mixture was heated to 80° C. until all the solids dissolved.
3. The solution was cooled to 60° C. and water (1 mL) was added slowly.

...

7. TGA mass loss was 10.5%.
8. Yield: 0.3 g.

Semantic Search

Q: What **solvents** are used in the reactions that produce **(2-methyl-4-(4-methyl-1-piperazinyl**...?

Patent ID	Solvent	Full Procedure
US06169084	ethyl acetate / toluene	A 0.5 g sample of...
...	...	...

SYNKB

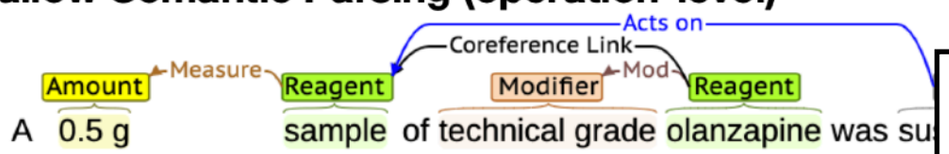
Q: What **amount** is **ethyl acetate** at when used to suspend **olanzapine**?

Patent ID	Amount	Full Procedure
US06169084	10 mL	A 0.5 g sample of...
...	...	...

SYNKB

Extracted Information

1. Shallow Semantic Parsing (operation-level)



2. Slot Filling (procedure-level)

Patent ID	Reagent	Solvent
US06169084	olanzapine	ethyl acetate / to
...	...	...

SYNKB: Semantic Search for Synthetic Procedures

Fan Bai[♣] **Alan Ritter**[♣] **Peter Madrid**[♠] **Dayne Freitag**[◇] **John Niekrasz**[◇]

♣ School of Interactive Computing, Georgia Institute of Technology

♠ Biosciences Division, SRI International

◇ Artificial Intelligence Center, SRI International

{fan.bai, alan.ritter}@cc.gatech.edu

{peter.madrid, daynefreitag, john.niekrasz}@sri.com

Demo URL: https://tinyurl.com/synkb

Example (SynKB)

*"What are the **solvents** used for reactions containing the reagent **triphosgene**?"*

Enter Your Search Query

Semantic Slot Search:

Product

Reagent: triphosgene

Solvent: ?

Other Compound

Starting Material

Temperature

Reaction Time

Yield Other

Yield Percent

Semantic Parse Search:

SYNKB: Semantic Search for Synthetic Procedures

Fan Bai[♣] **Alan Ritter**[♣] **Peter Madrid**[♠] **Dayne Freitag**[◇] **John Niekrasz**[◇]
♣ School of Interactive Computing, Georgia Institute of Technology
♠ Biosciences Division, SRI International
◇ Artificial Intelligence Center, SRI International
{fan.bai, alan.ritter}@cc.gatech.edu
{peter.madrid, daynefreitag, john.niekrasz}@sri.com
Demo URL: https://tinyurl.com/synkb

Example (SynKB)

*"What are the **solvents** used for reactions containing the reagent **triphosgene**?"*

Enter Your Search Query

Semantic Slot Search:

Product

Reagent: triphosgene

Solvent: ?

Other Com

Reaction T

Semantic Par

Search Results

reagent : triethylamine \ Triphosgene
solvent : Chloroform
count: 71

reagent : triphosgene \ triethylamine \ triethylamine
solvent : chloroform \ chloroform \ chloroform
count: 11

reagent : HCl \ triphosgene \ triethylamine
solvent : dioxane \ DCM
count: 9

reagent : triphosgene \ Et3N \ Et3N
solvent : CH2Cl2 \ CH2Cl2(4 \ CH2Cl2
count: 8

reagent : triphosgene \ triethylamine
solvent : toluene
count: 7

reagent : triphosgene
solvent : toluene

SynKB:
60 solvents
(vs Reaxys' 8)

SYNKB: Semantic Search for Synthetic Procedures

Fan Bai[♣] **Alan Ritter**[♣] **Peter Madrid**[♠] **Dayne Freitag**[◇] **John Niekrasz**[◇]
♣ School of Interactive Computing, Georgia Institute of Technology
♠ Biosciences Division, SRI International
◇ Artificial Intelligence Center, SRI International
{fan.bai, alan.ritter}@cc.gatech.edu
{peter.madrid, daynefreitag, john.niekrasz}@sri.com
Demo URL: https://tinyurl.com/synkb

Example (SynKB)

*"What are the **solvents** used for reactions containing the reagent **triphosgene**?"*

Enter Your Search Query

Semantic Slot Search:

Product

Reagent: triphosgene

Solvent: ?

Other Com

Reaction T

Search Results

reagent : triethylamine \ Triphosgene
solvent : Chloroform
count: 71

reagent : triphosgene \ triethylamine \ triethylamine
solvent : chloroform \ chloroform \ chloroform

reagent : triphosgene
solvent : toluene

SynKB:
60 solvents
(vs Reaxys' 8)

Peter Madrid
@Molecreationist

Replying to @loadingfan

Fan, this is a great free tool that my whole group is using now and will continue to get better. Great collaboration.

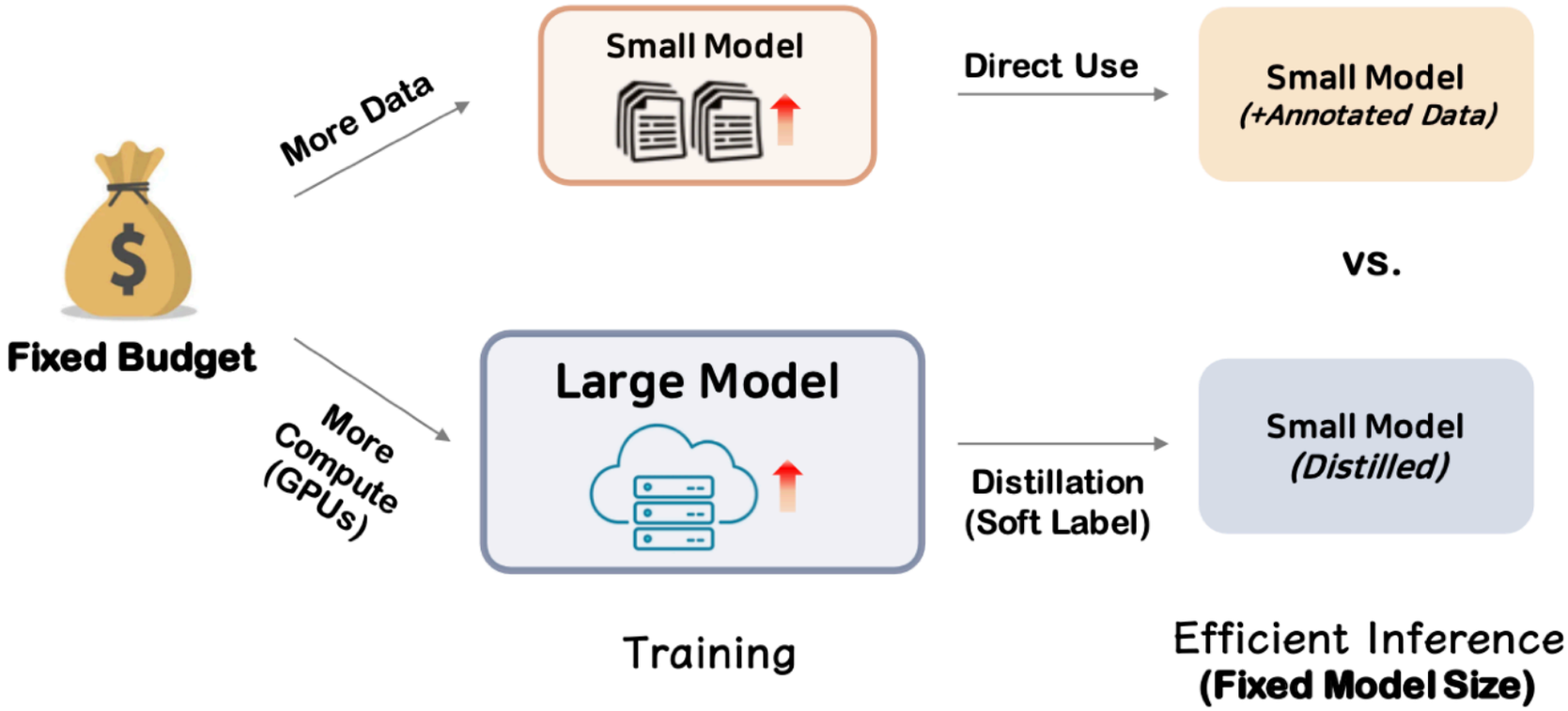
8:29 PM · Oct 8, 2022 · Twitter for iPhone

SYNKB: Semantic Search for Synthetic Procedures

Fan Bai[♣] **Alan Ritter**[♣] **Peter Madrid**[♠] **Dayne Freitag**[◇] **John Niekrasz**[◇]
♣ School of Interactive Computing, Georgia Institute of Technology
♠ Biosciences Division, SRI International
◇ Artificial Intelligence Center, SRI International
{fan.bai, alan.ritter}@cc.gatech.edu
{peter.madrid, daynefreitag, john.niekrasz}@sri.com
Demo URL: https://tinyurl.com/synkb

Computation vs. Annotation

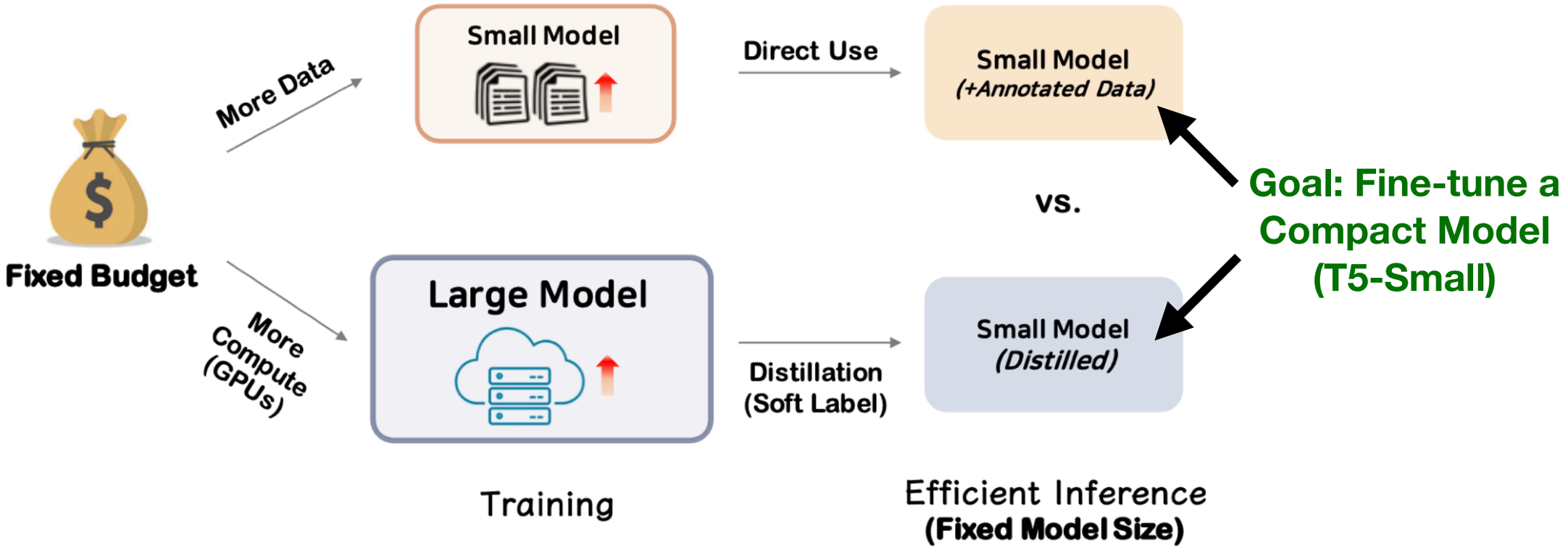
Tradeoff #2: Distill or Annotate?



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Computation vs. Annotation

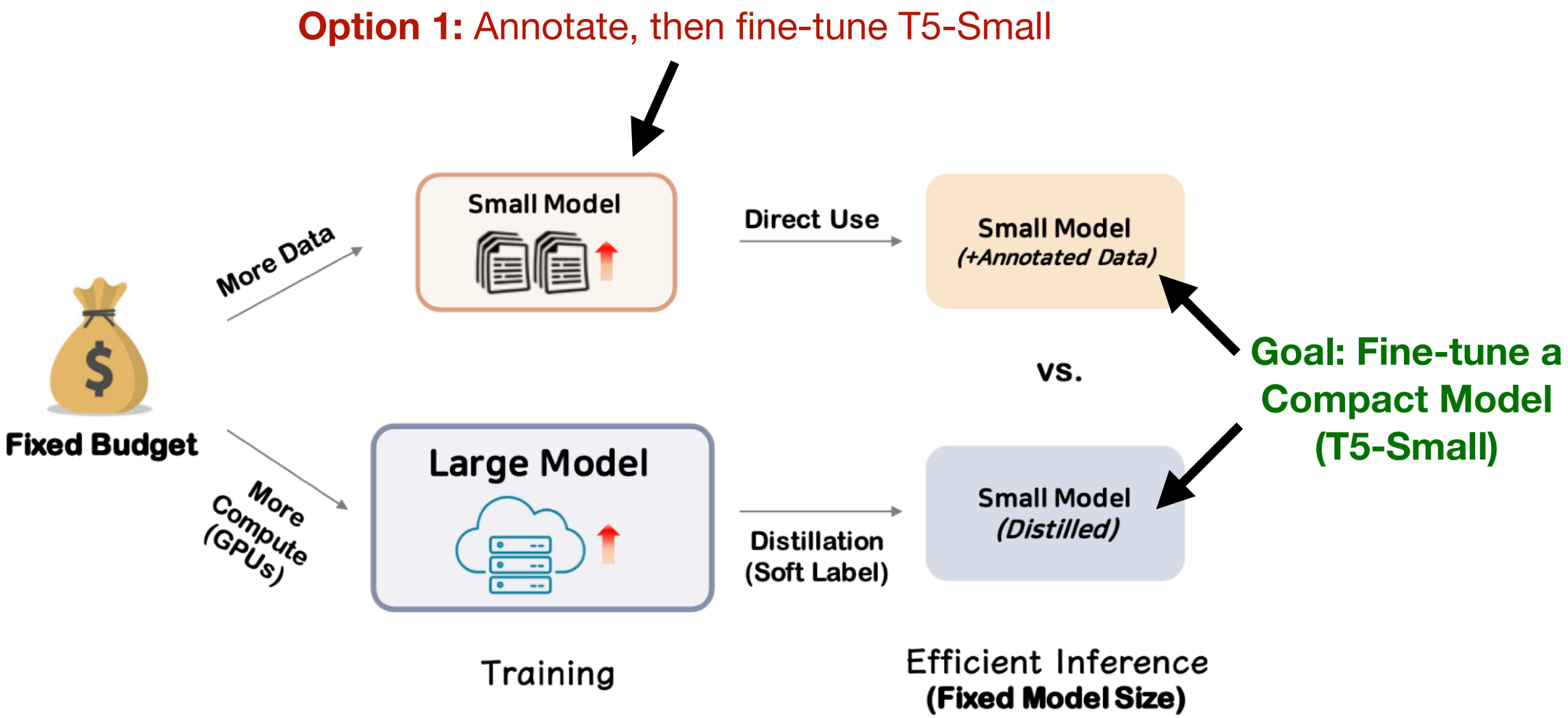
Tradeoff #2: Distill or Annotate?



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Computation vs. Annotation

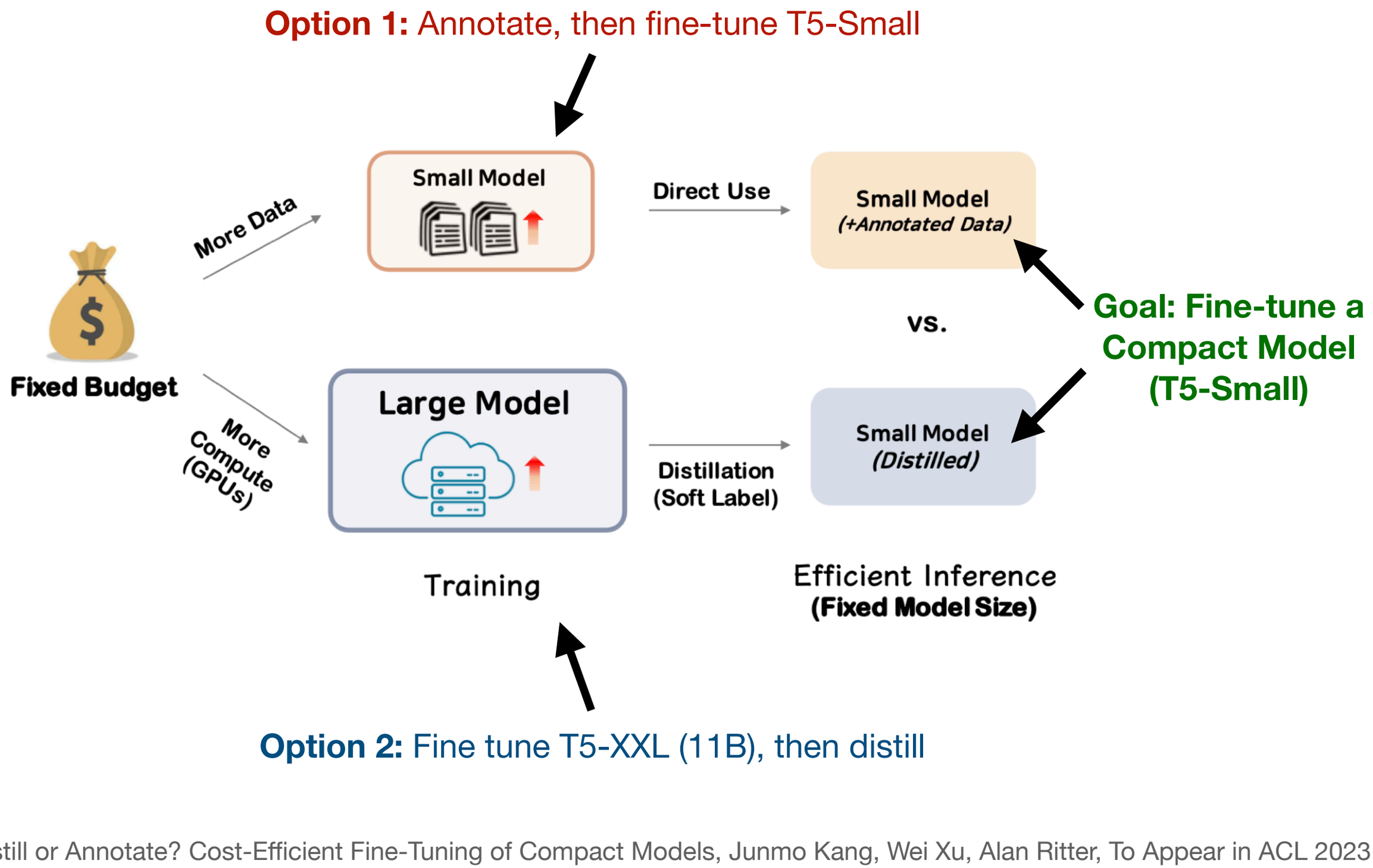
Tradeoff #2: Distill or Annotate?



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

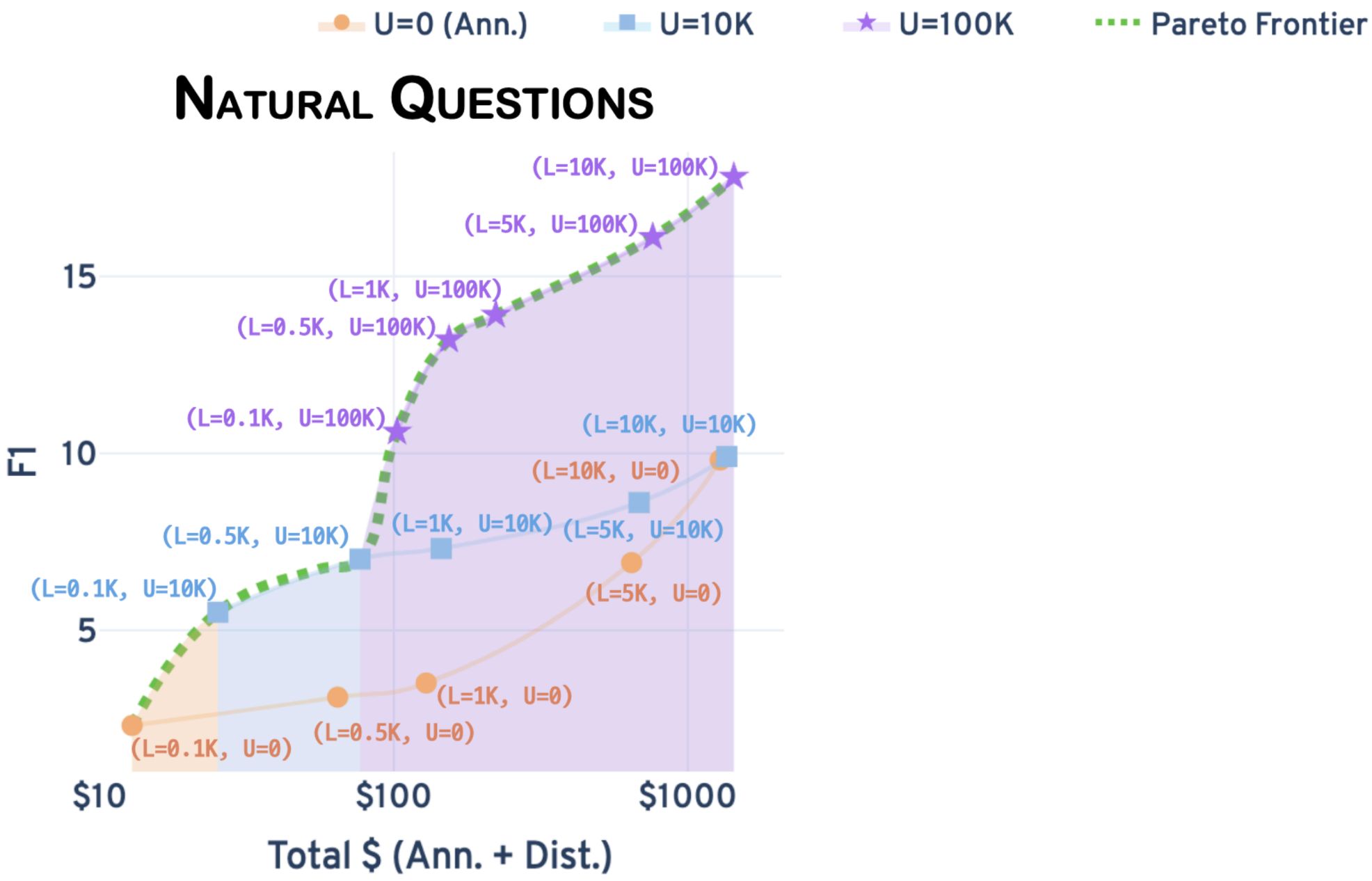
Computation vs. Annotation

Tradeoff #2: Distill or Annotate?



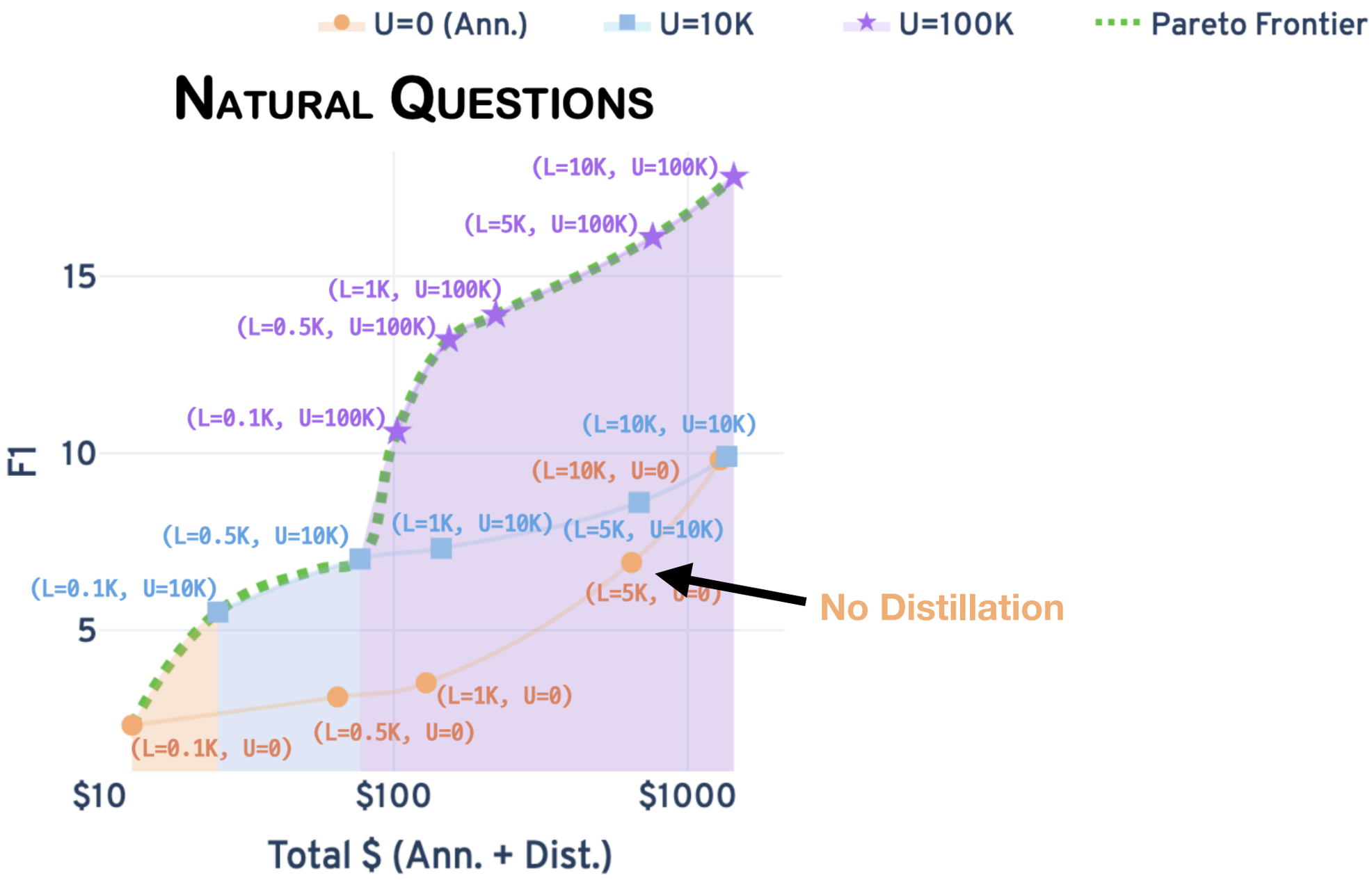
Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Pareto Curves



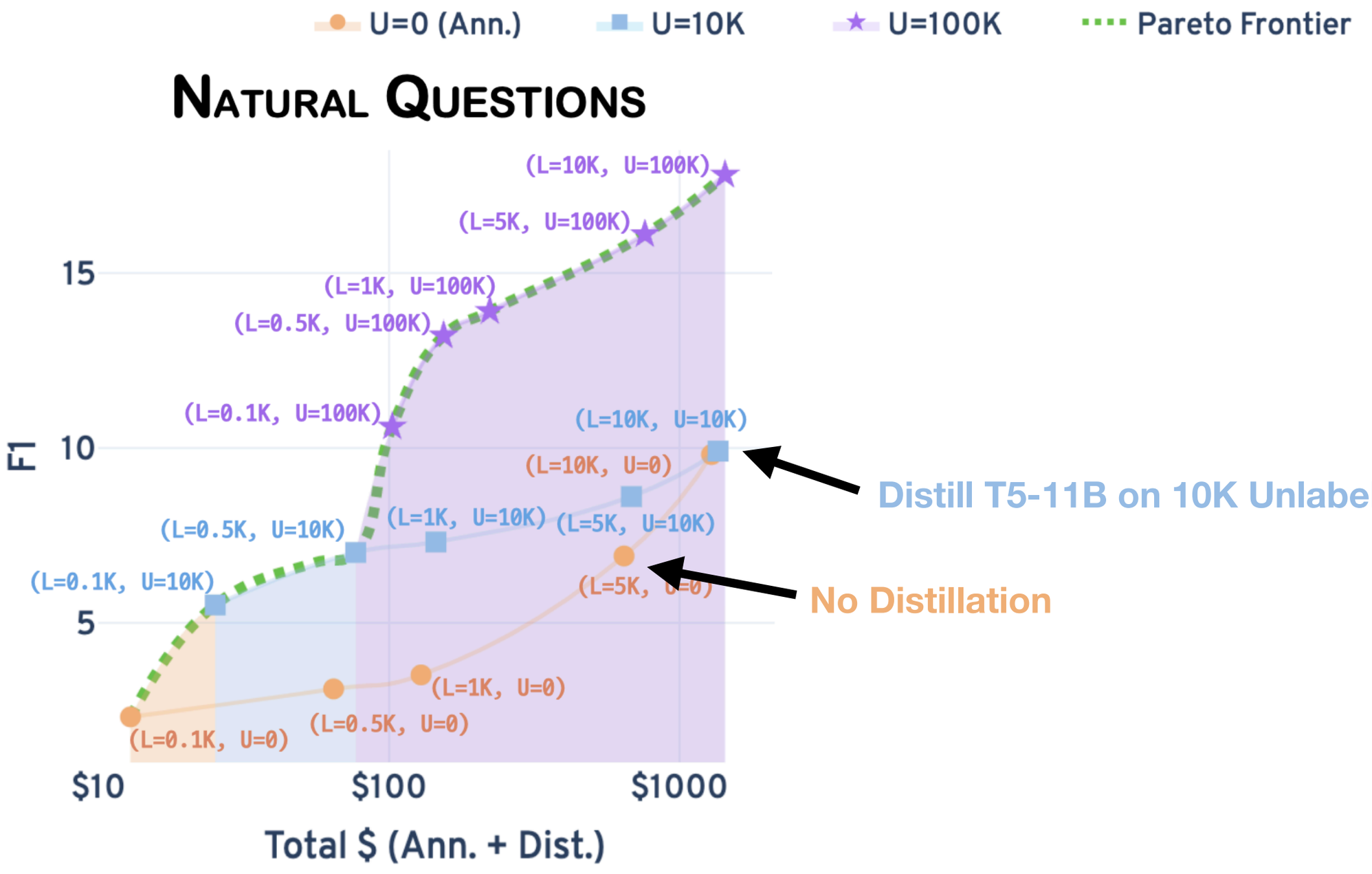
Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Pareto Curves



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Pareto Curves

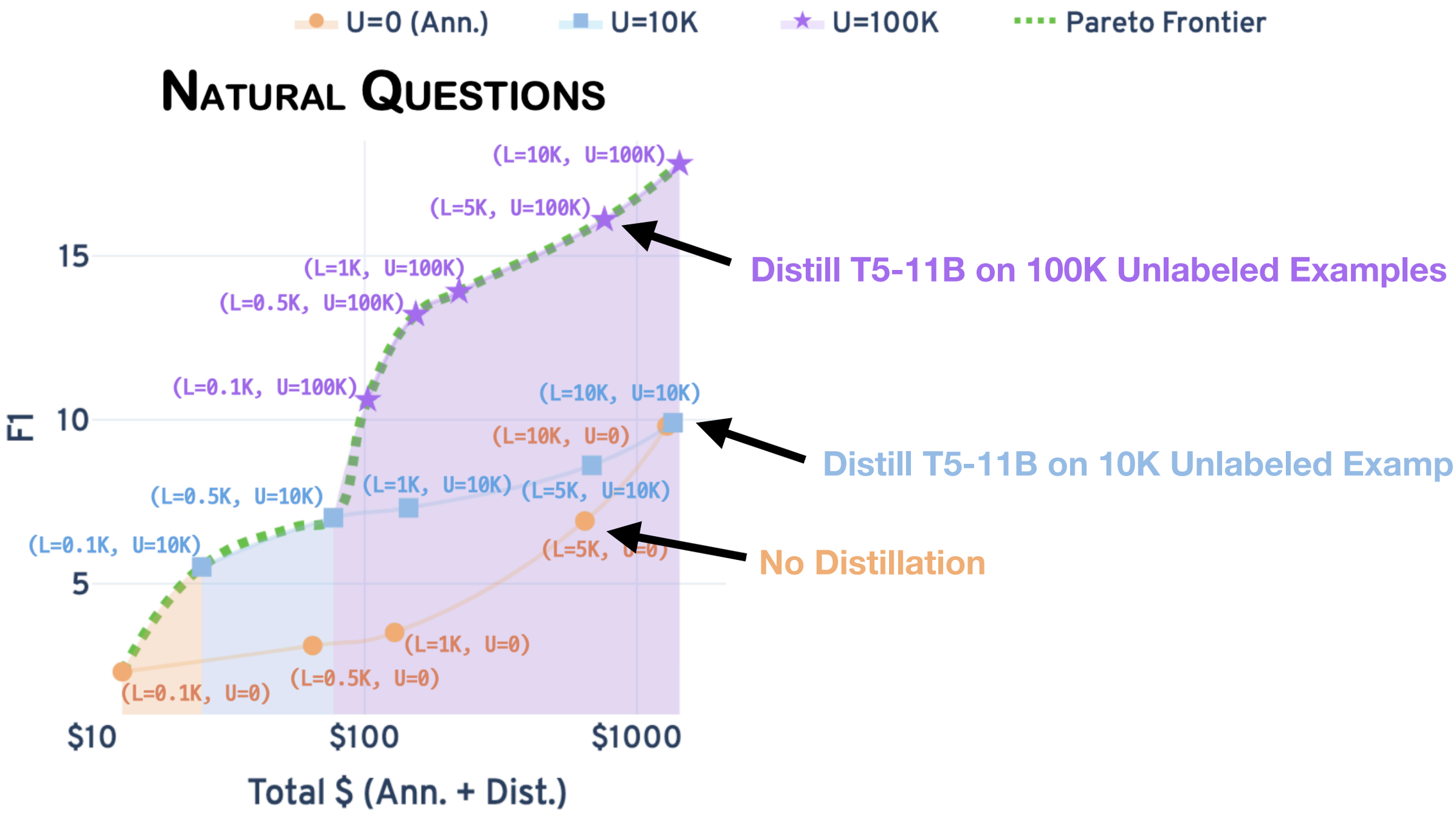


Distill T5-11B on 10K Unlabeled Examples

No Distillation

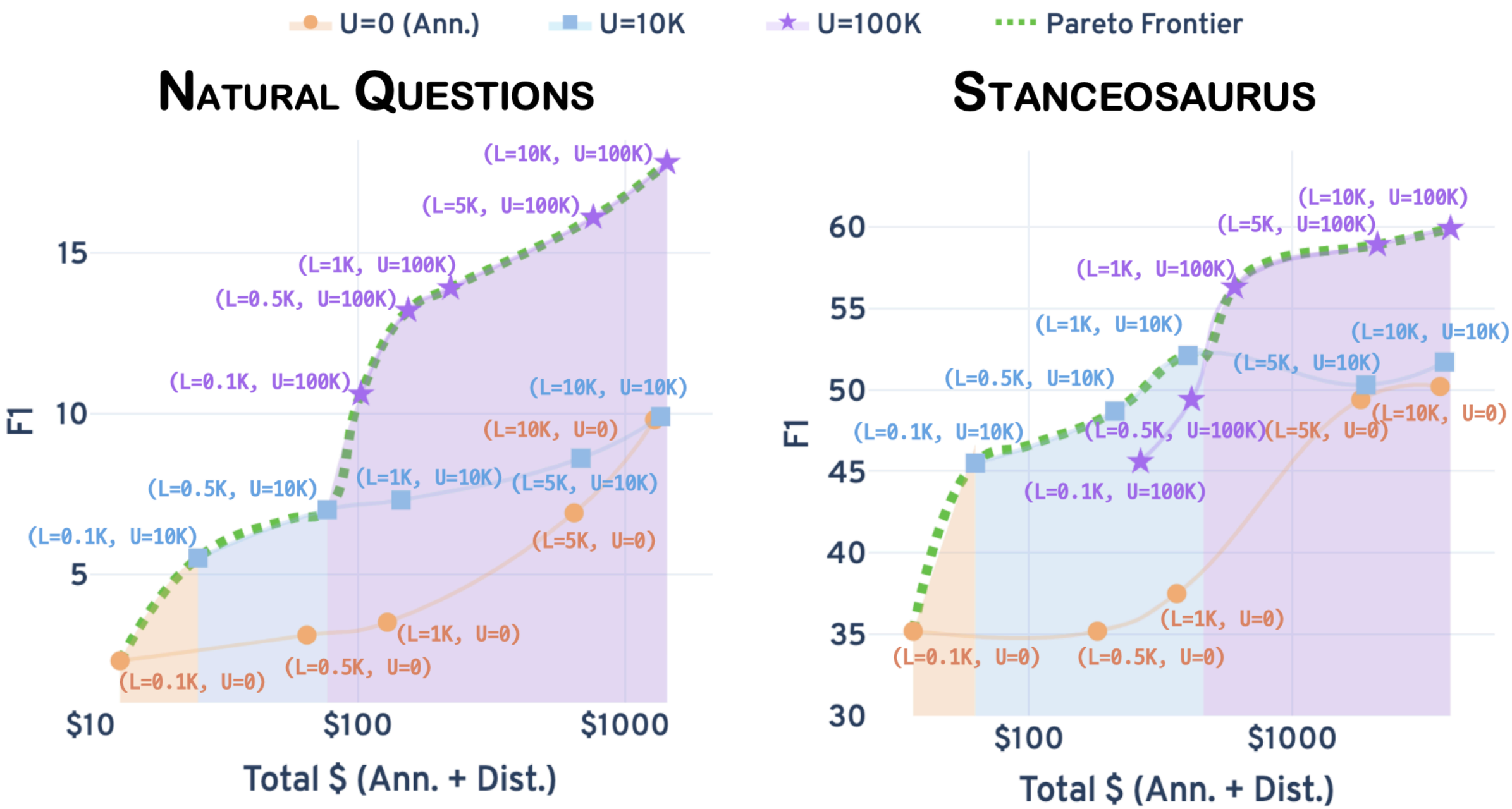
Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Pareto Curves



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Pareto Curves

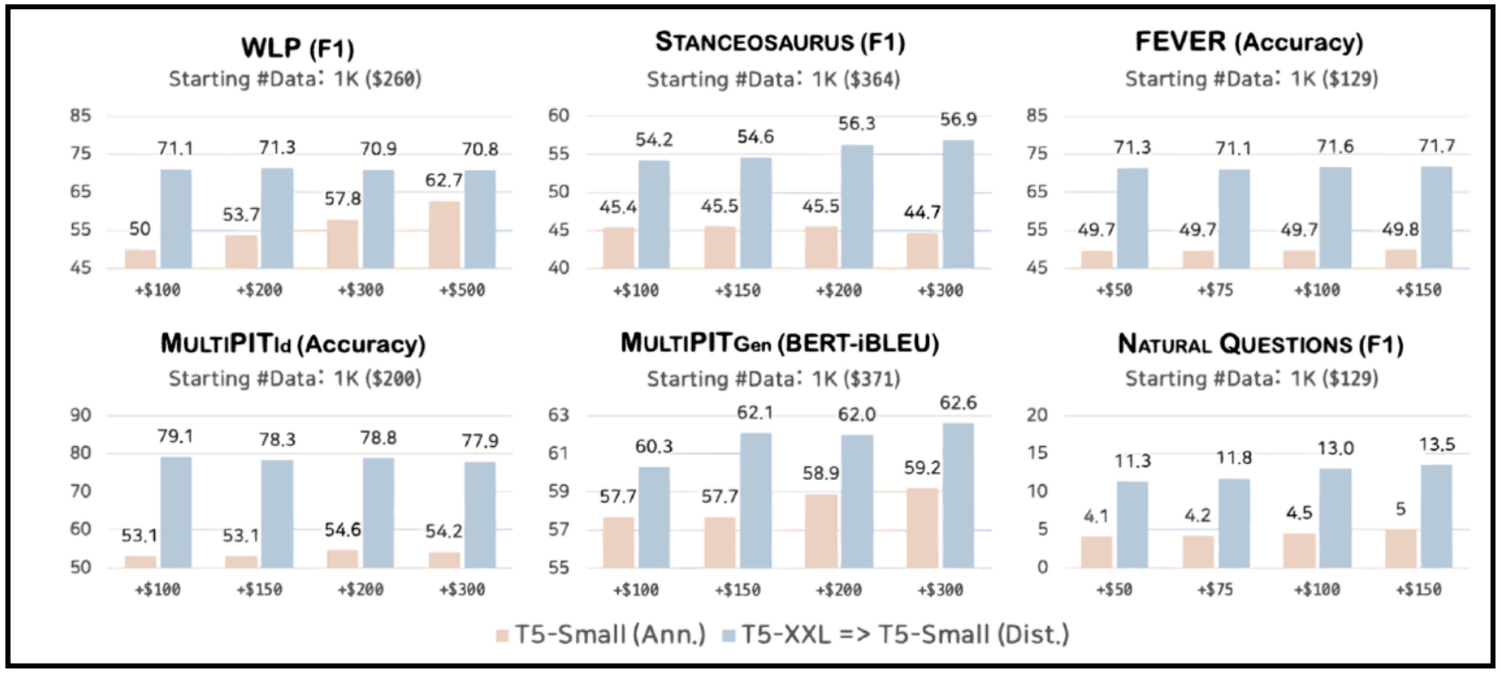


Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

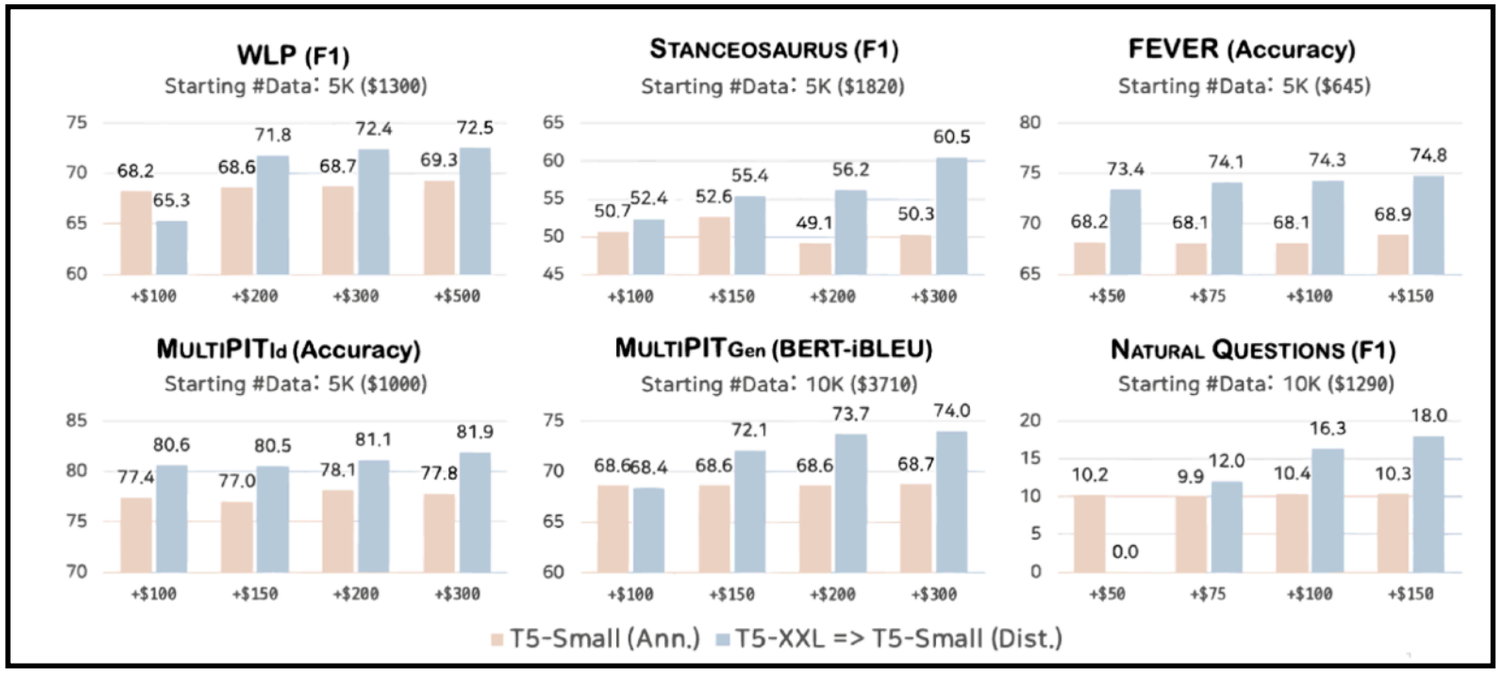
Distill or Annotate?

- **Distillation is usually more economical than annotation.**

Small Initial Training Dataset

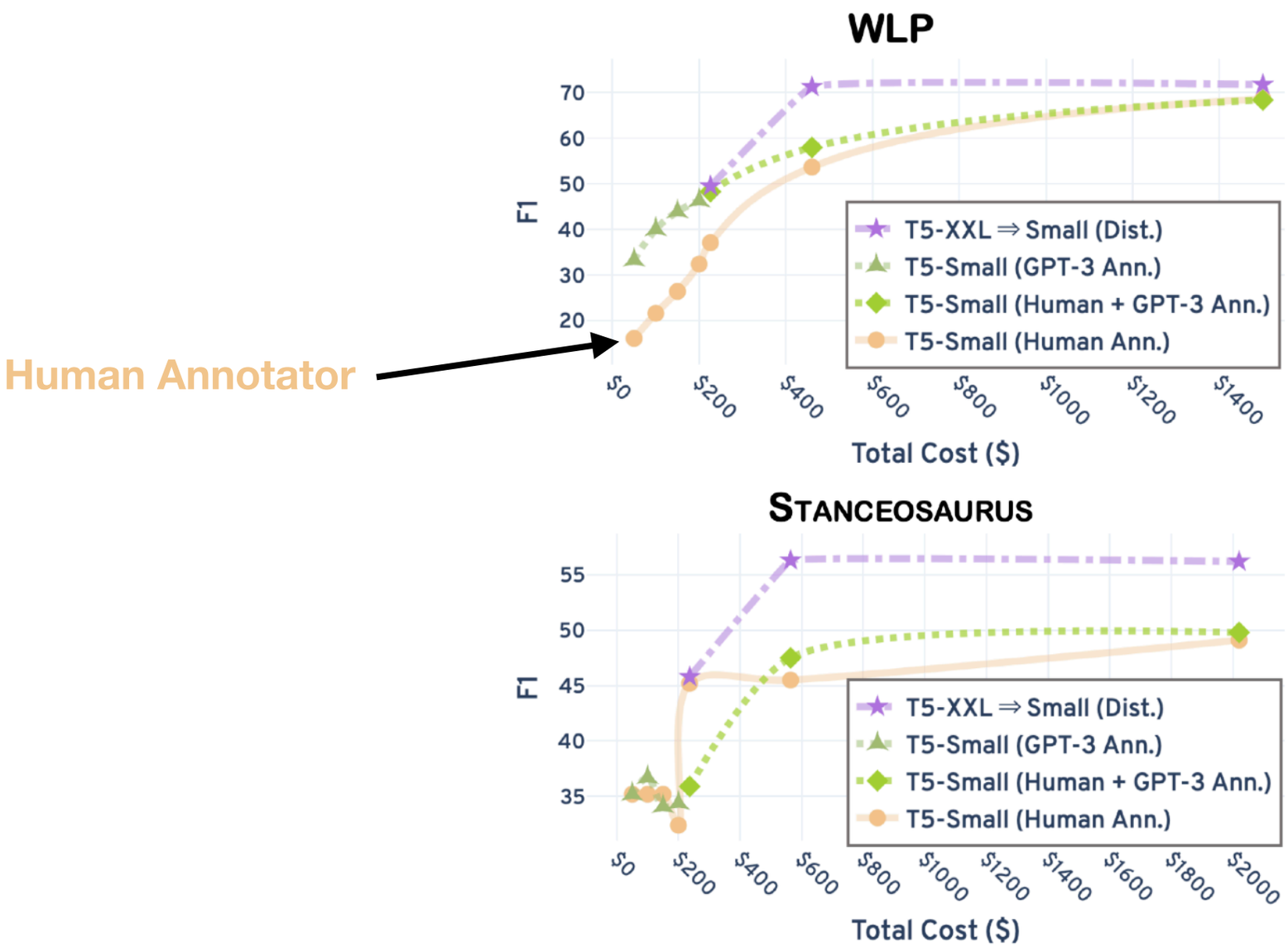


Large Initial Training Dataset



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

GPT-3.5 as an Annotator?



Human Annotator

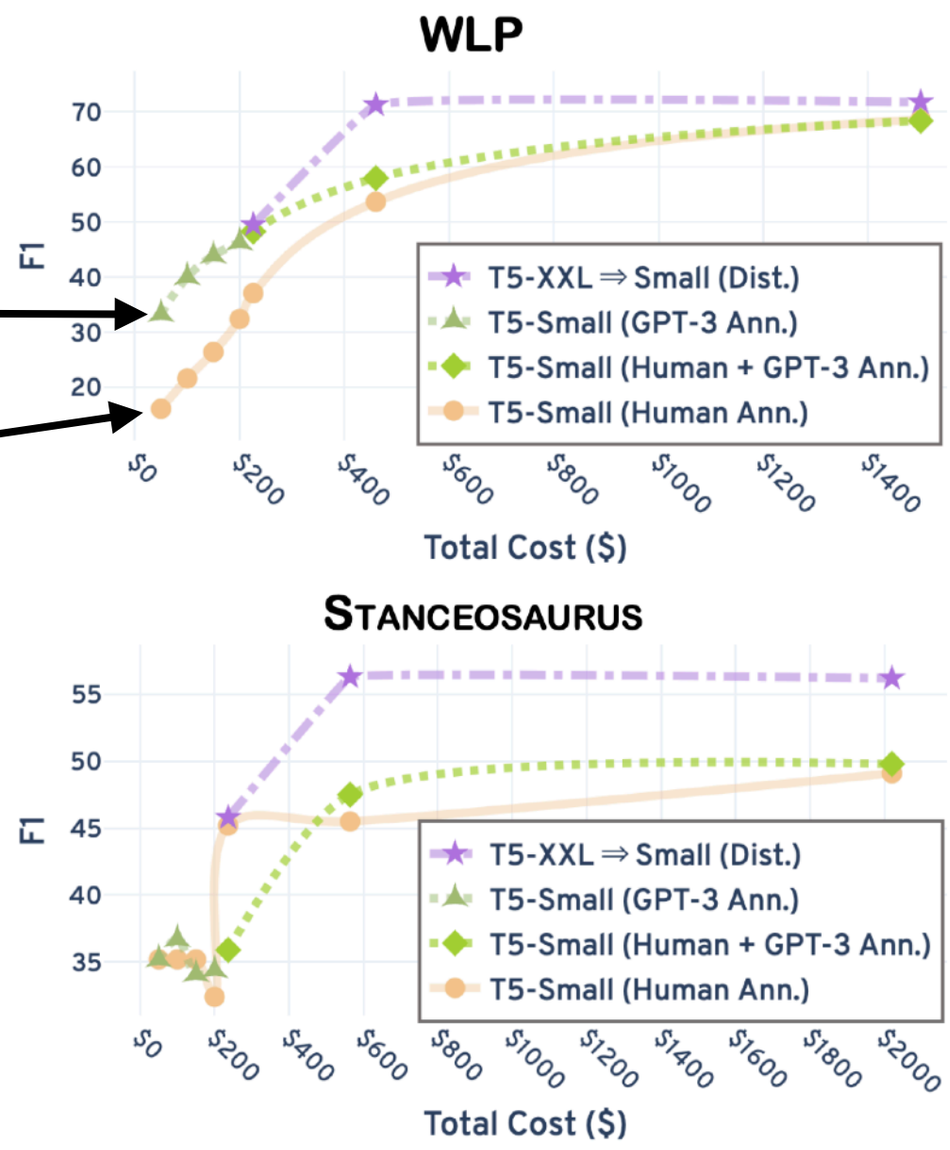
Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Related Work: *Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. EMNLP Findings 2021*

GPT-3.5 as an Annotator?

GPT-3.5 Annotator

Human Annotator



Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

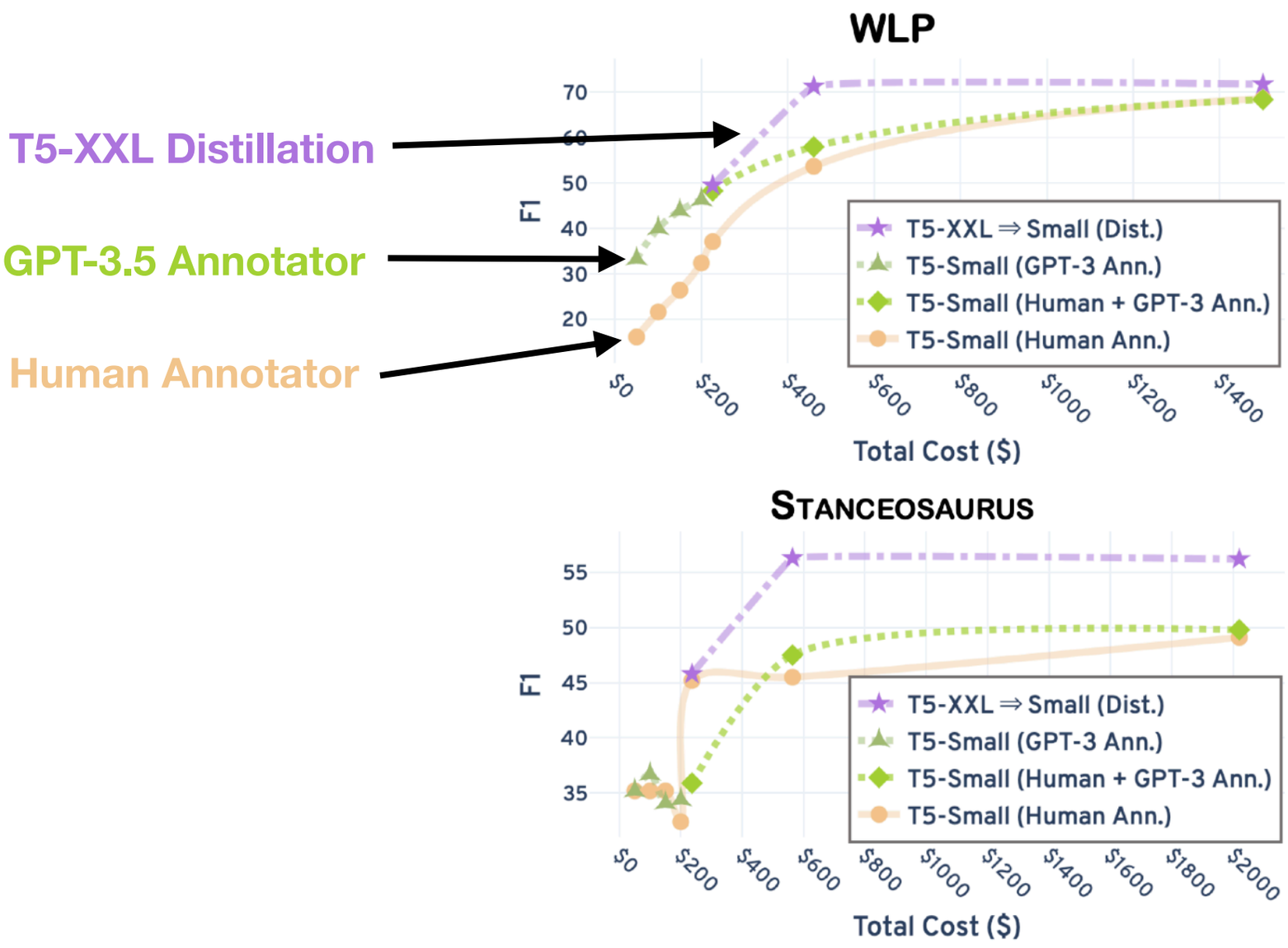
Related Work: *Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. EMNLP Findings 2021*

GPT-3.5 as an Annotator?

T5-XXL Distillation

GPT-3.5 Annotator

Human Annotator

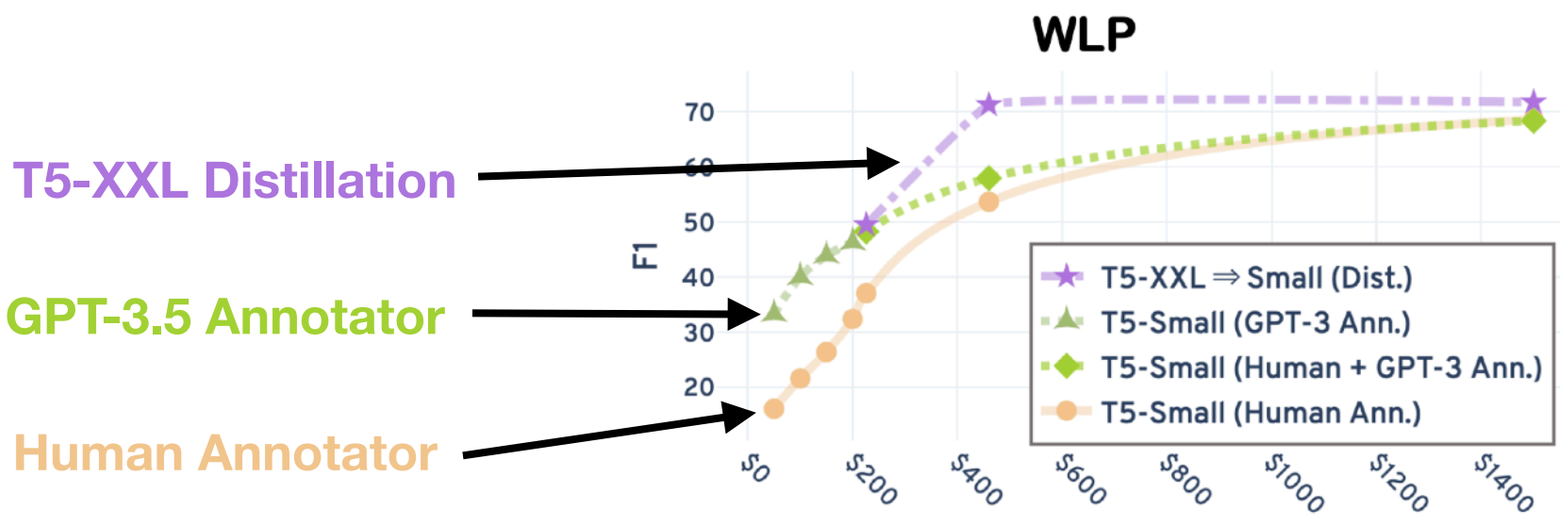


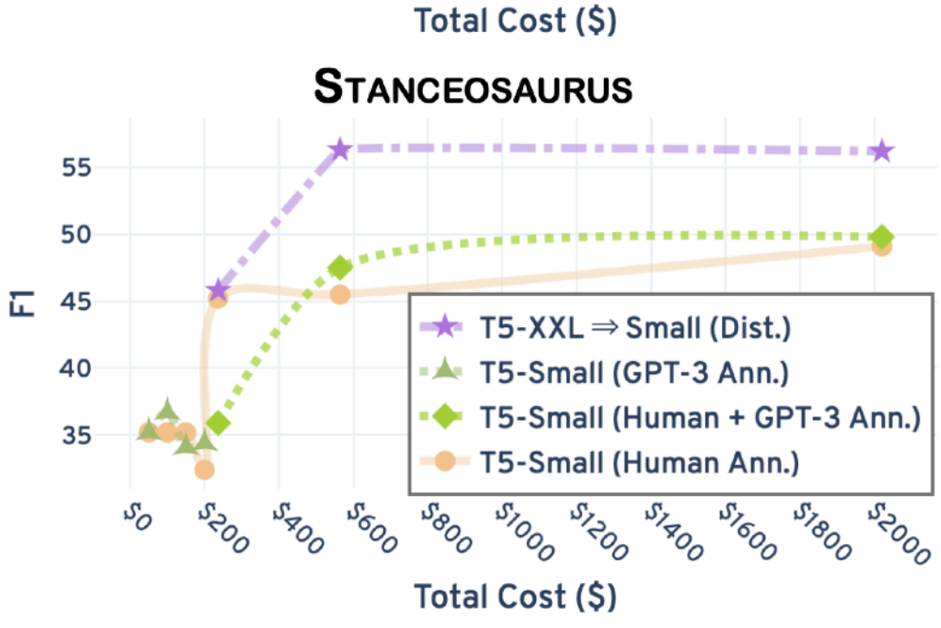
Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Related Work: *Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. EMNLP Findings 2021*

GPT-3.5 as an Annotator?

- Distilling fine-tuned T5-XXL is more economical



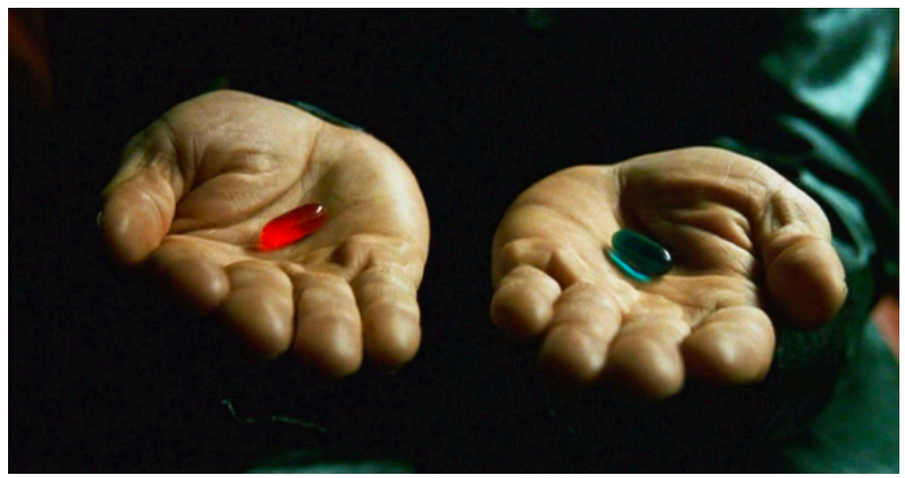


Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Junmo Kang, Wei Xu, Alan Ritter, To Appear in ACL 2023

Related Work: *Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. EMNLP Findings 2021*

Economical Use of Pre-trained Models

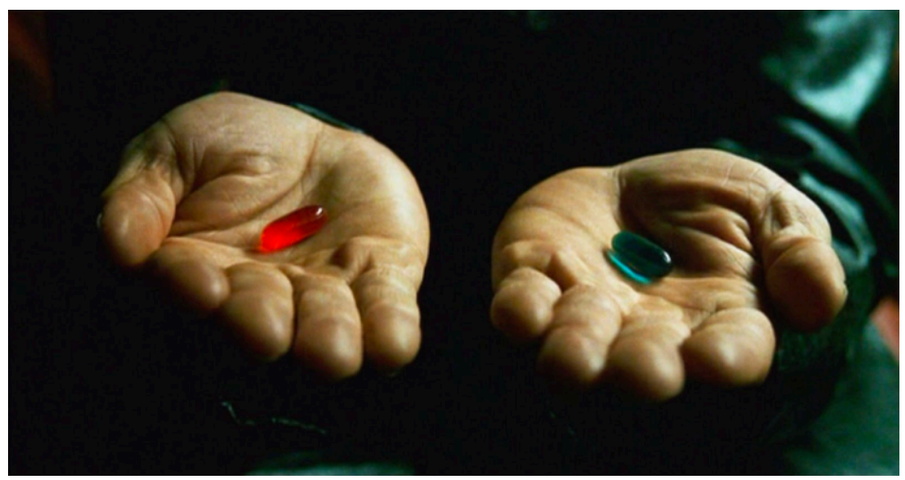
- **So far**: Computation vs. Annotation
  - Adapting to a new domain
  - Knowledge Distillation



Annotation **Computation**

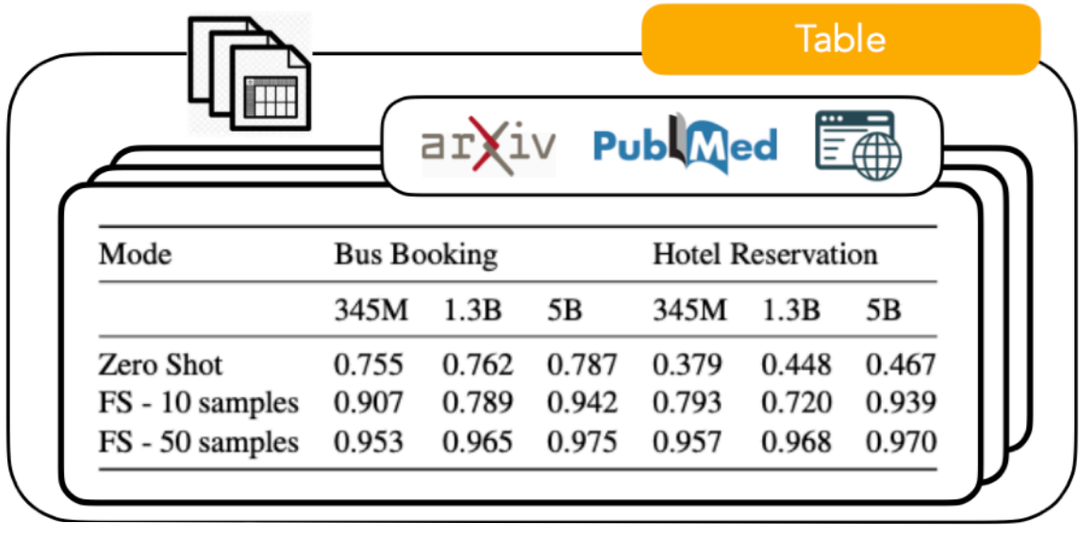
Economical Use of Pre-trained Models

- **So far**: Computation vs. Annotation
  - Adapting to a new domain
  - Knowledge Distillation

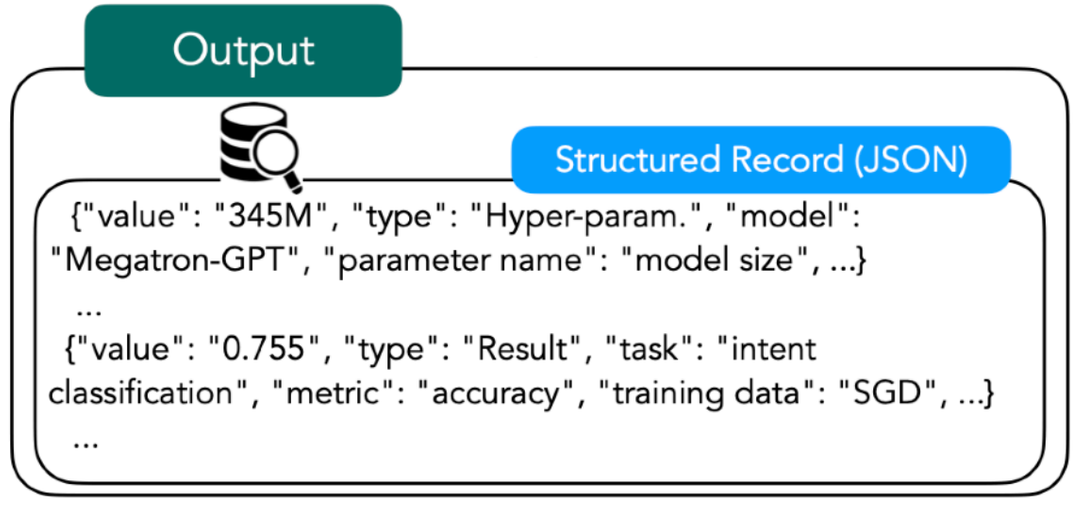


Annotation **Computation**

- **Next**: Cost-Efficient Data Extraction from Tables



Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970



Extracting Data from Tables

- Extracting Leaderboards from **ML Papers**, e.g. AxCell (Kadres et. al. 2020)
- **Materials Science**, e.g. DiSCoMaT (Gupta et. al. 2023)
- **Chemistry** Tables, e.g. ChemDataExtractor (Swain et. Al. 2016)
- **Webpages** e.g. OpenCeres (Lockard et. al. 2019, 2020)

	Dataset / Method	OPIEC59k				ReVerb45k			
		Macro F1	**Micro F1**	**Pair F1**	***Avg***	**Macro F1**	**Micro F1**	**Pair F1**	***Avg***
	Optimal Clust.	80.3					93.5	92.1	90.1
	CMVC	52.8					87.9	89.4	81.1
	KMeans	53.5 ±0.0				0.0	89.1 ±0.0	89.3 ±0.0	82.7
ours	PCKMeans	58.7 ±0.0	91.5 ±0.0	86.1 ±0.0	78.7	72.0 ±0.0	88.5 ±0.0	87.0 ±0.0	82.5
ours	LLM Correction	58.7	91.5	85.2	78.4	69.9	89.2	88.4	82.5
ours	Keyphrase Clust.	**60.3** ±0.0	**92.5** ±0.0	**87.3** ±0.0	**80.0**	**72.3** ±0.0	**90.2** ±0.0	**90.0** ±0.0	**84.2**

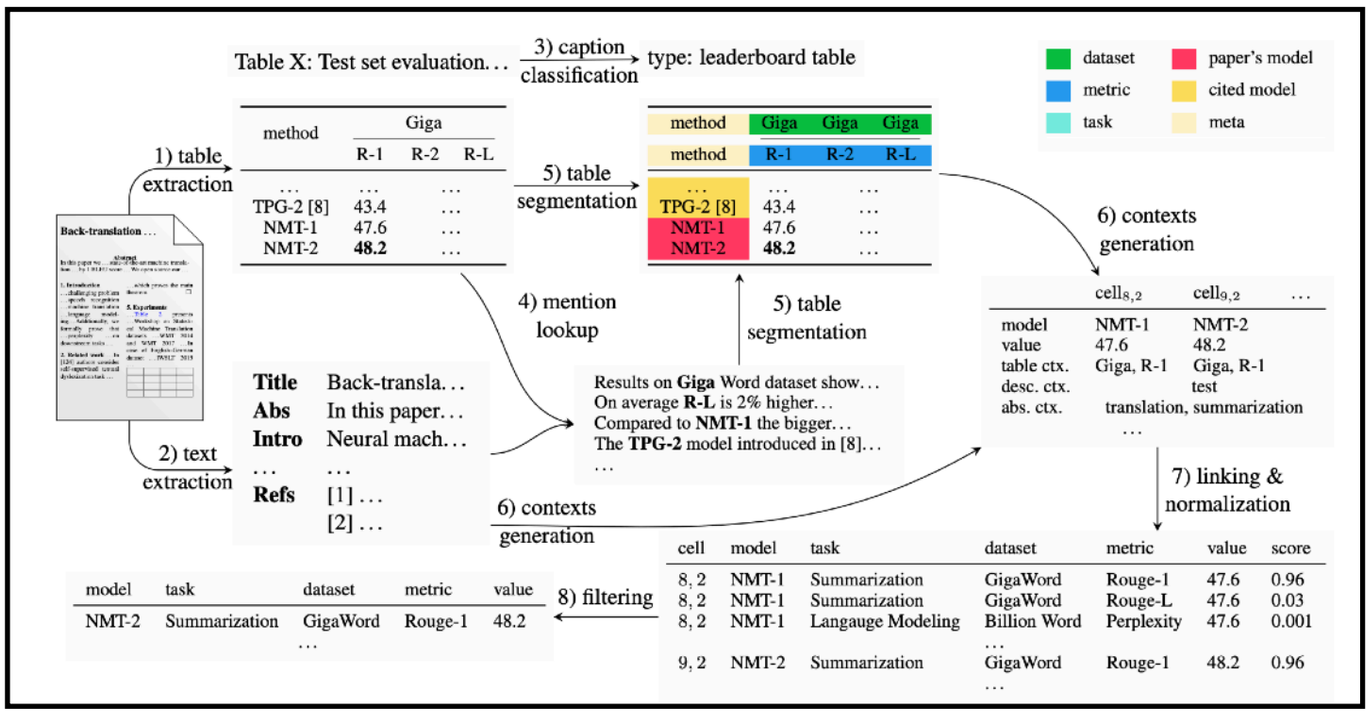
Sample ML Table

(Viswanathan et. al. 2023)

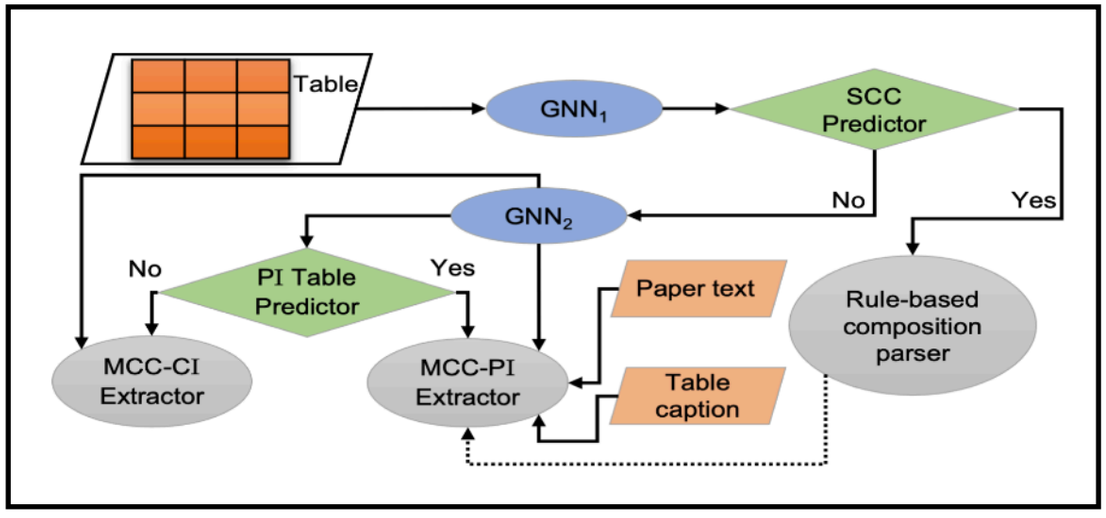
Extracting Data from Tables

- Extracting Leaderboards from **ML Papers**, e.g. AxCell (Kadres et. al. 2020)
- **Materials Science**, e.g. DiSCoMaT (Gupta et. al. 2023)
- **Chemistry** Tables, e.g. ChemDataExtractor (Swain et. Al. 2016)
- **Webpages** e.g. OpenCeres (Lockard et. al. 2019, 2020)

AxCell (ML Leaderboards)



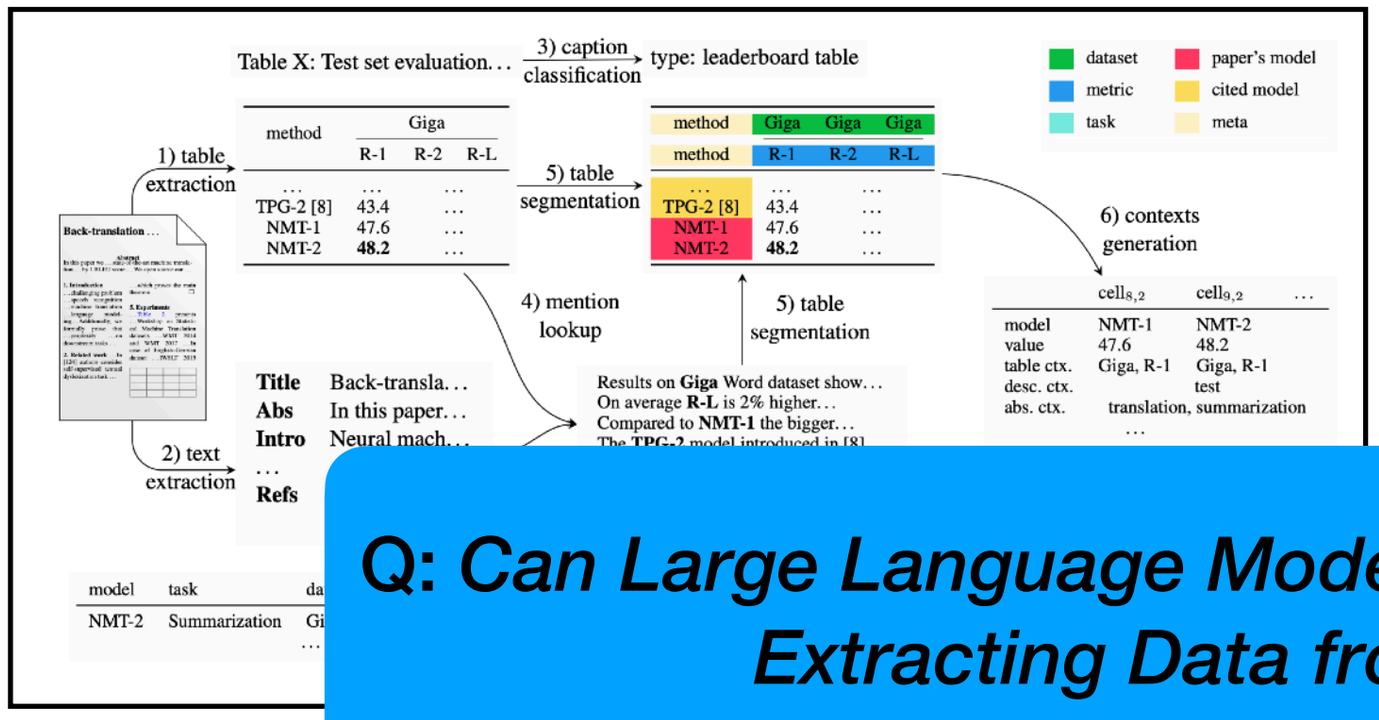
DiSCoMaT (Materials Science)



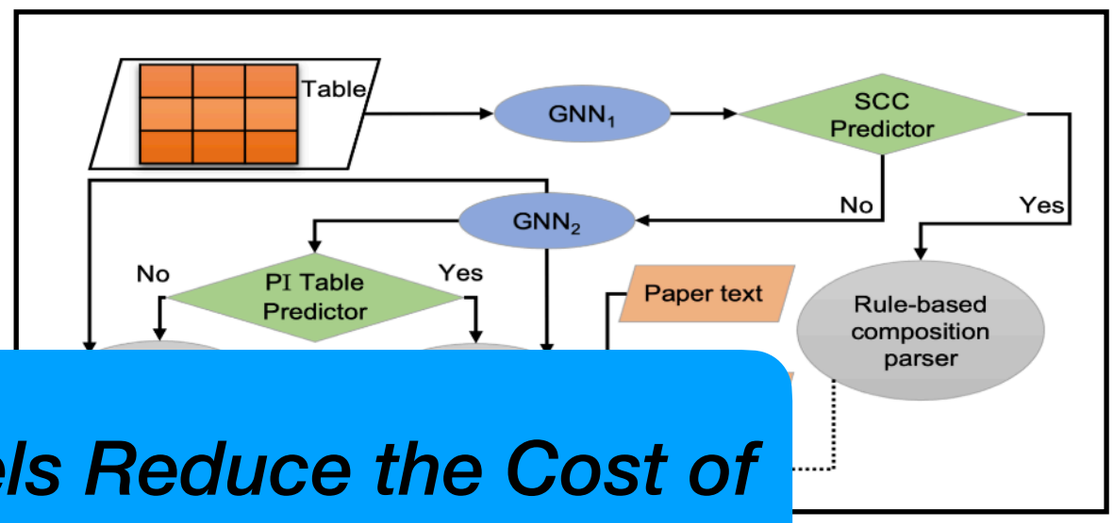
Extracting Data from Tables

- Extracting Leaderboards from **ML Papers**, e.g. AxCell (Kadres et. al. 2020)
- **Materials Science**, e.g. DiSCoMaT (Gupta et. al. 2023)
- **Chemistry** Tables, e.g. ChemDataExtractor (Swain et. Al. 2016)
- **Webpages** e.g. OpenCeres (Lockard et. al. 2019, 2020)

AxCell (ML Leaderboards)



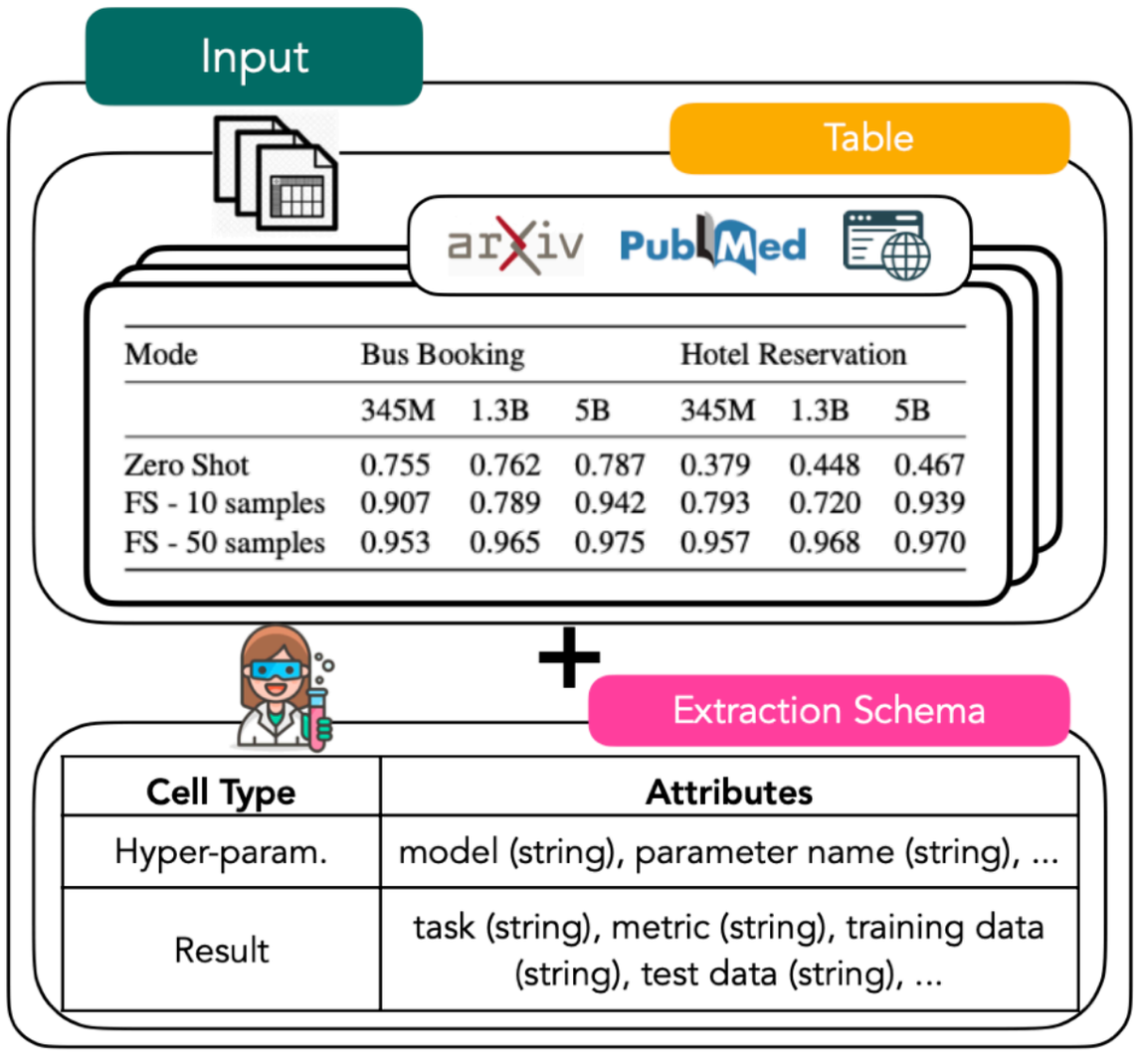
DiSCoMaT (Materials Science)



Q: *Can Large Language Models Reduce the Cost of Extracting Data from Tables?*

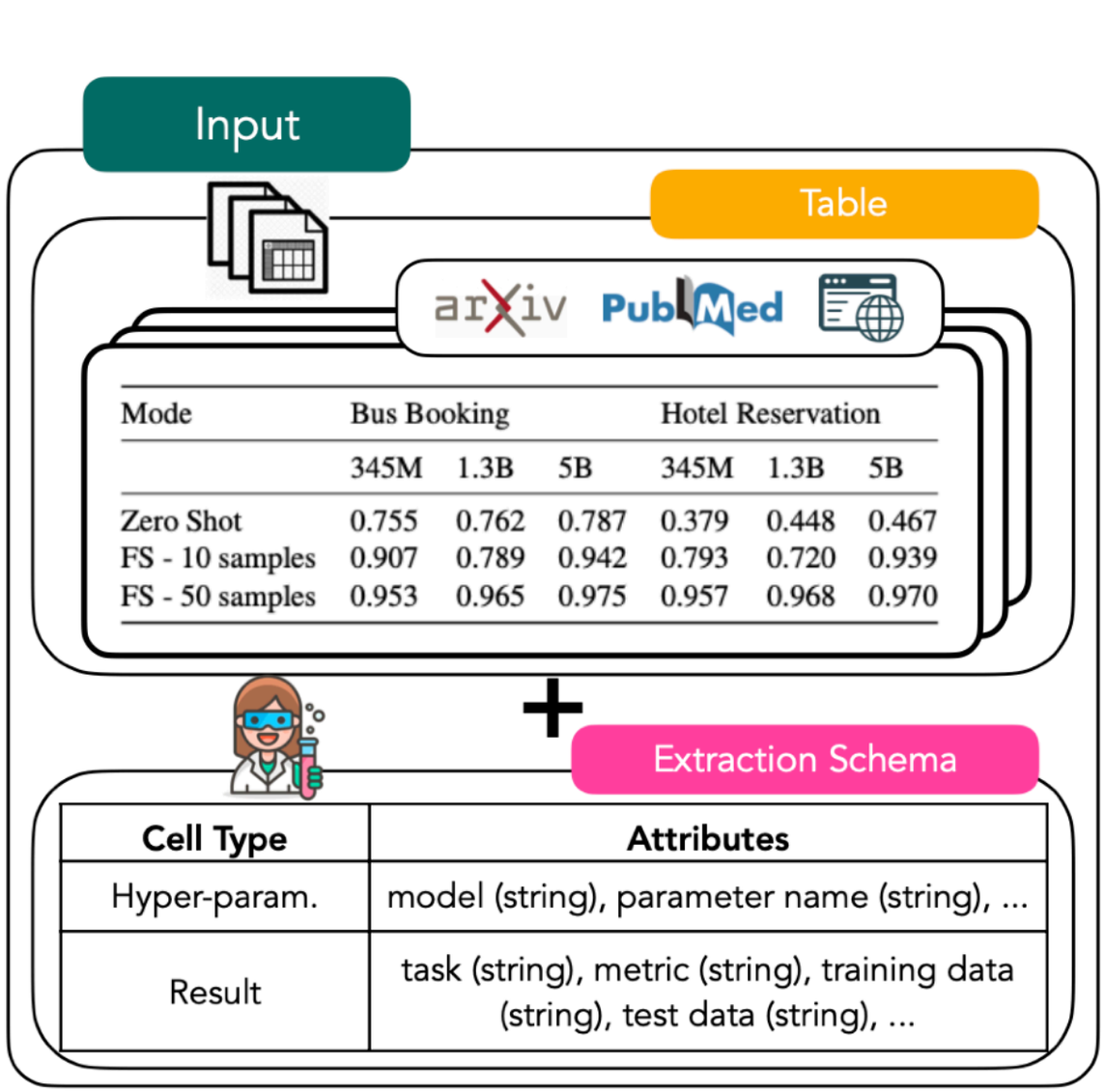
Schema-Driven Information Extraction

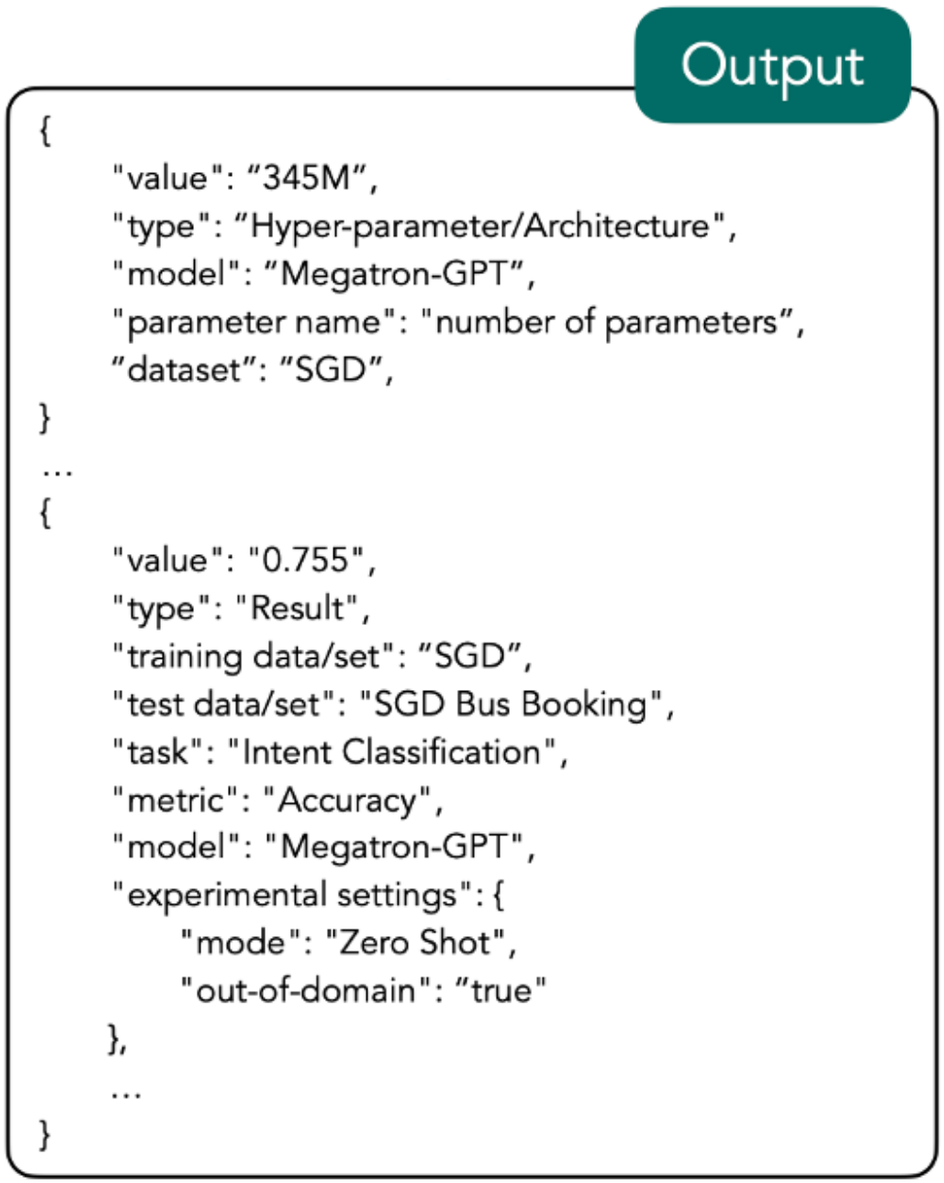
- Only supervision is an **Extraction Schema** authored by domain expert.



Schema-Driven Information Extraction

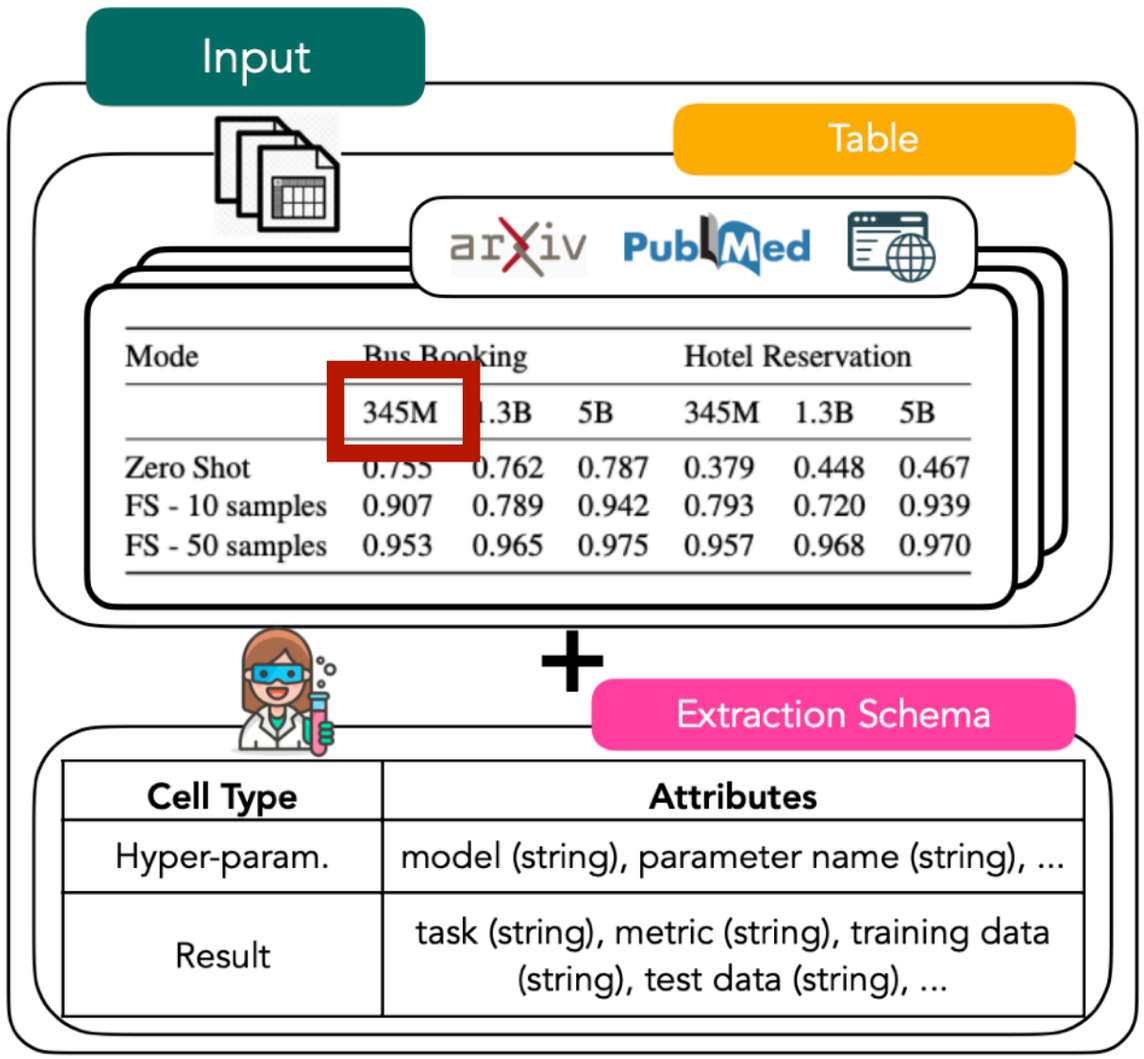
- Only supervision is an **Extraction Schema** authored by domain expert.

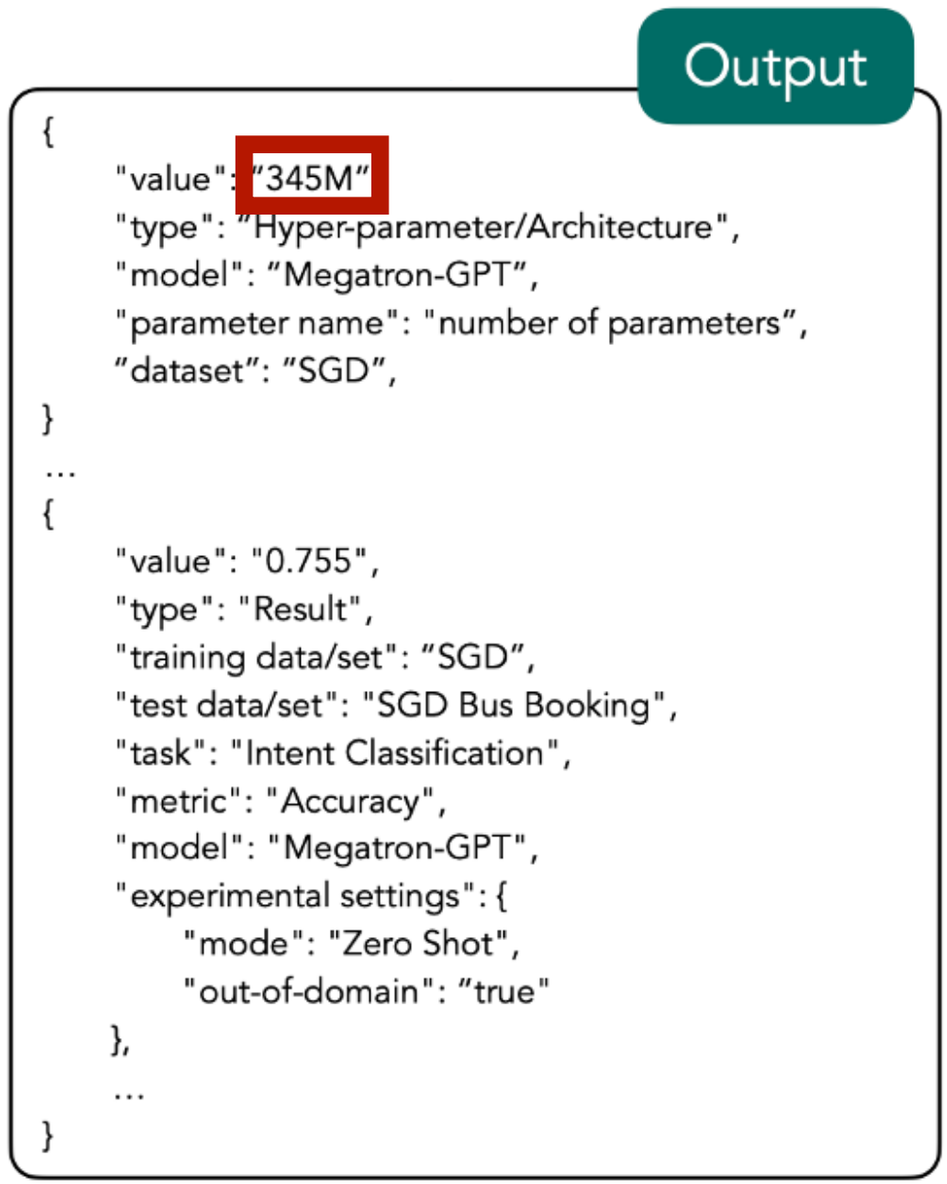




Schema-Driven Information Extraction

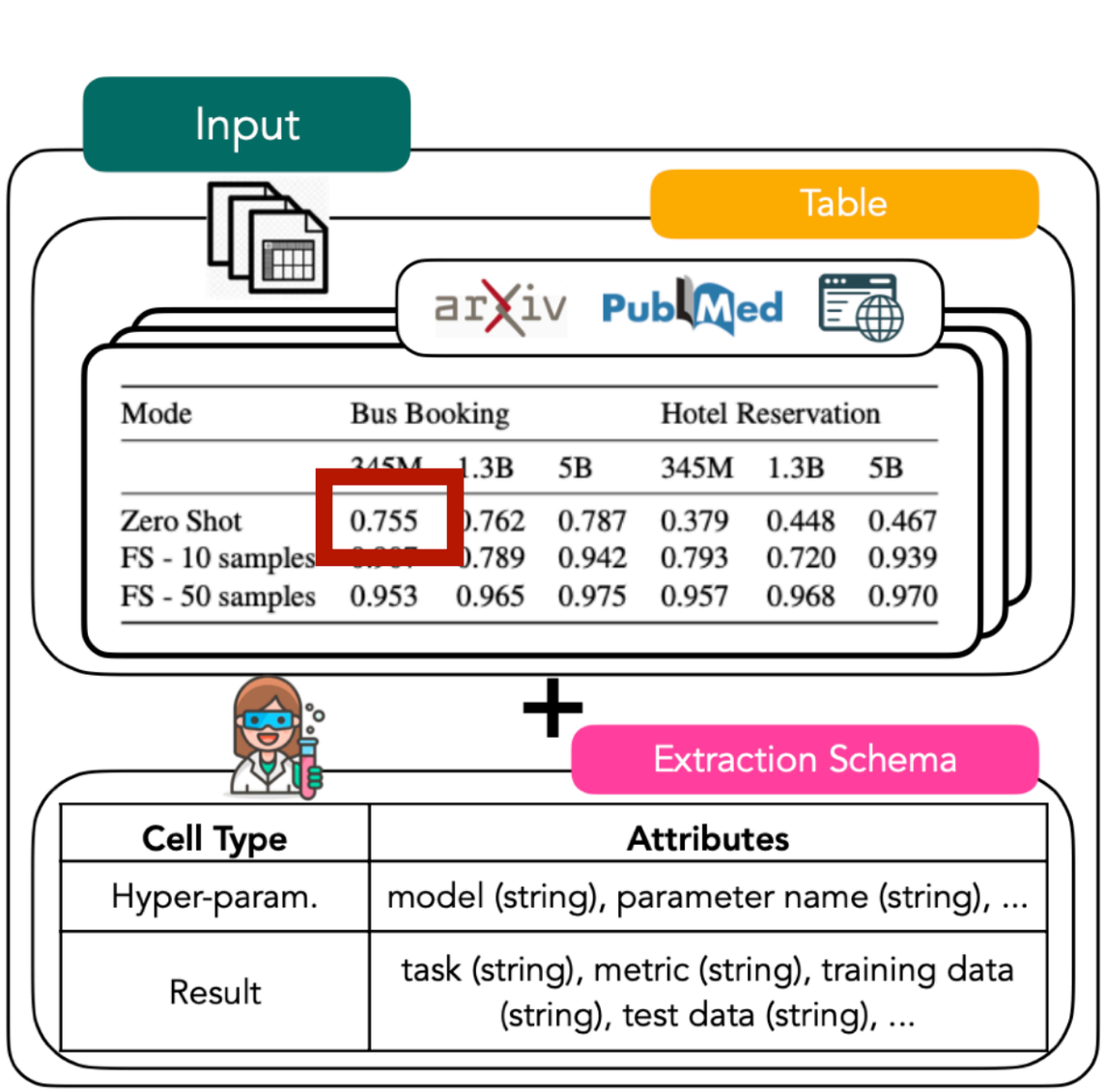
- Only supervision is an **Extraction Schema** authored by domain expert.

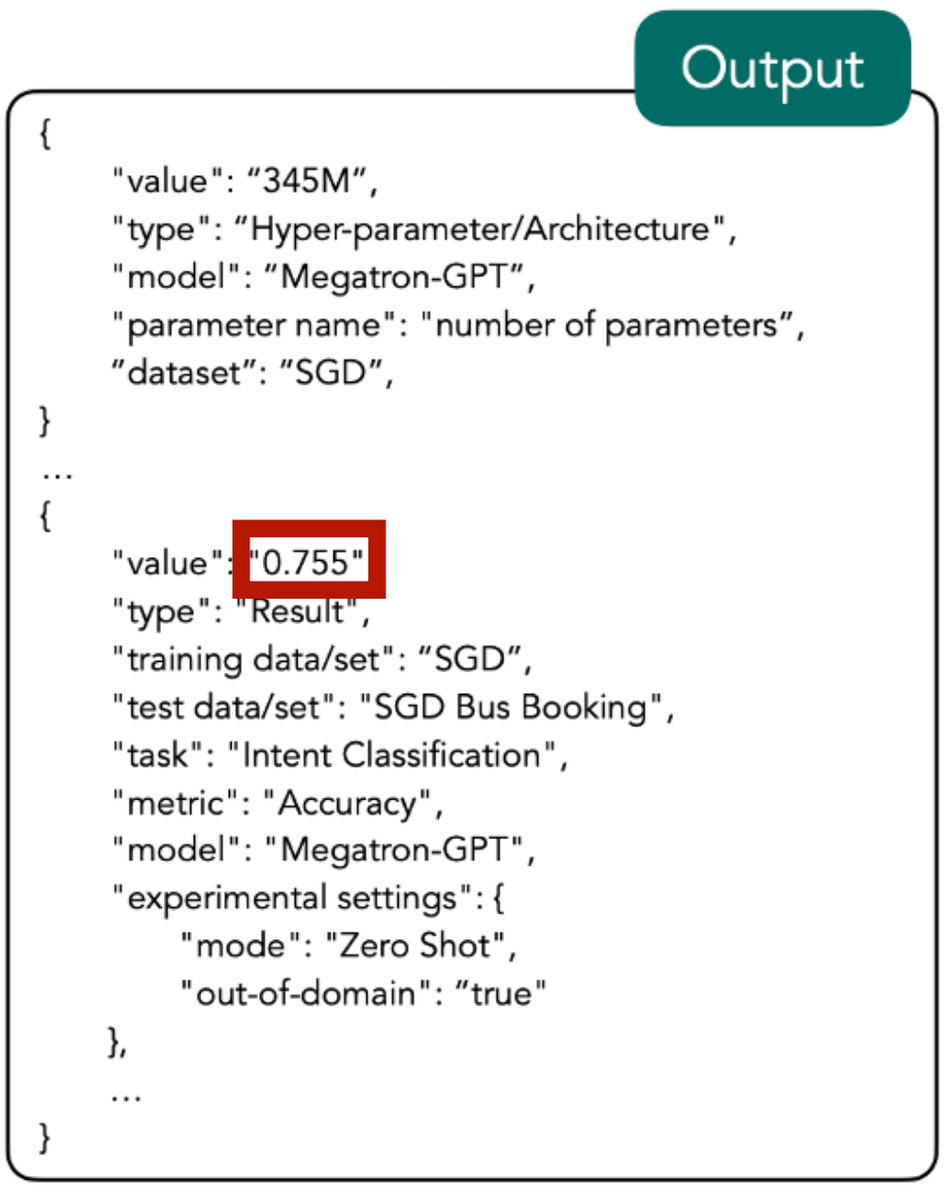




Schema-Driven Information Extraction

- Only supervision is an **Extraction Schema** authored by domain expert.





Benchmarking Schema-Driven IE

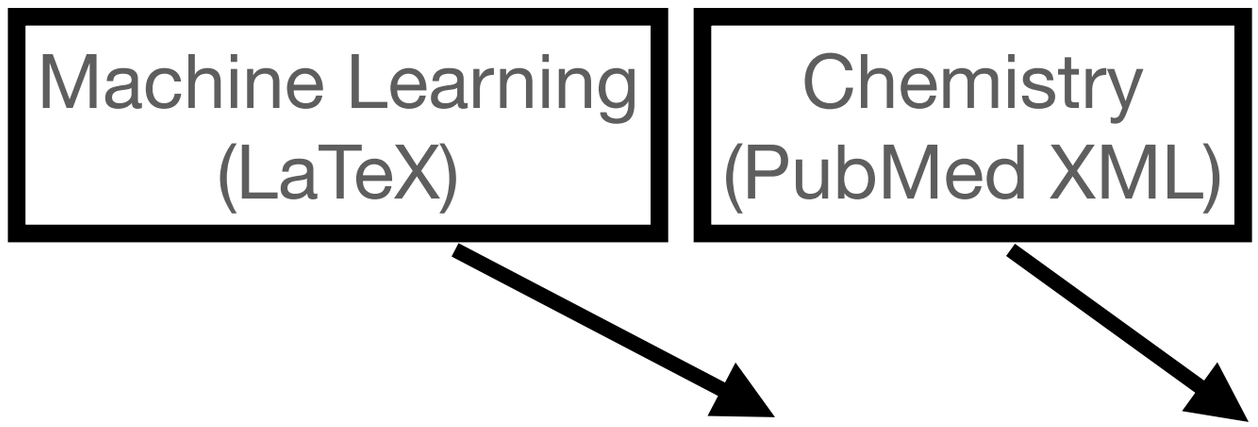
	MLTAB. (ours)	CHEMTAB. (ours)	DISCOMAT (Gupta et. al. 2023)	SWDE (Hao et. al. 2011)
# cell types	4	6	2	8
# papers (web.)	25	16	2,536	80
# tables (pages)	122	26	5,883	1,600
# anno. records	3,792	1,498	58,481	1,600

Benchmarking Schema-Driven IE

Machine Learning
(LaTeX)

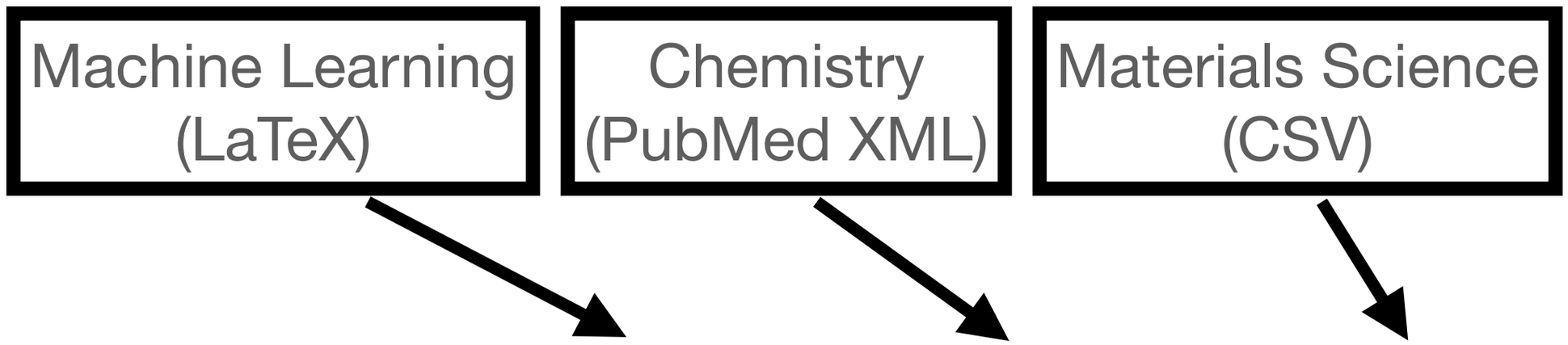
	MLTAB. (ours)	CHEMTAB. (ours)	DISCOMAT (Gupta et. al. 2023)	SWDE (Hao et. al. 2011)
# cell types	4	6	2	8
# papers (web.)	25	16	2,536	80
# tables (pages)	122	26	5,883	1,600
# anno. records	3,792	1,498	58,481	1,600

Benchmarking Schema-Driven IE



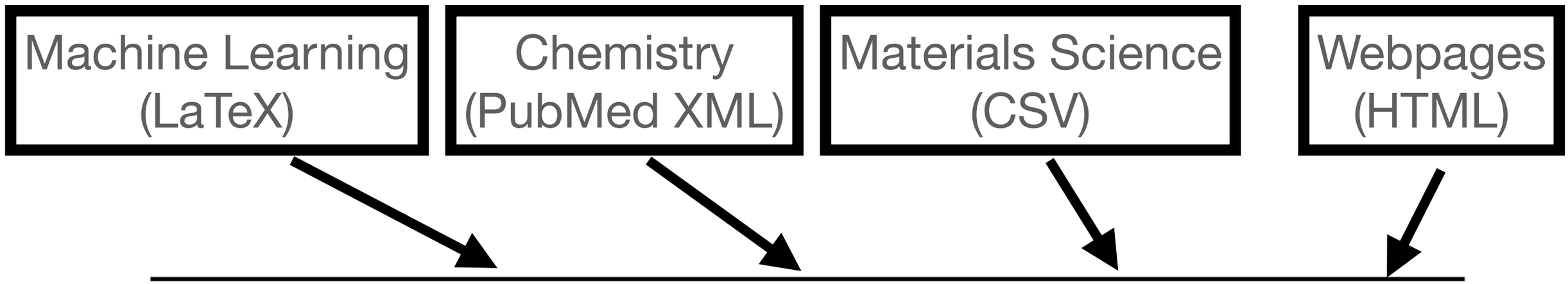
	MLTAB. (ours)	CHEMTAB. (ours)	DISCOMAT (Gupta et. al. 2023)	SWDE (Hao et. al. 2011)
# cell types	4	6	2	8
# papers (web.)	25	16	2,536	80
# tables (pages)	122	26	5,883	1,600
# anno. records	3,792	1,498	58,481	1,600

Benchmarking Schema-Driven IE



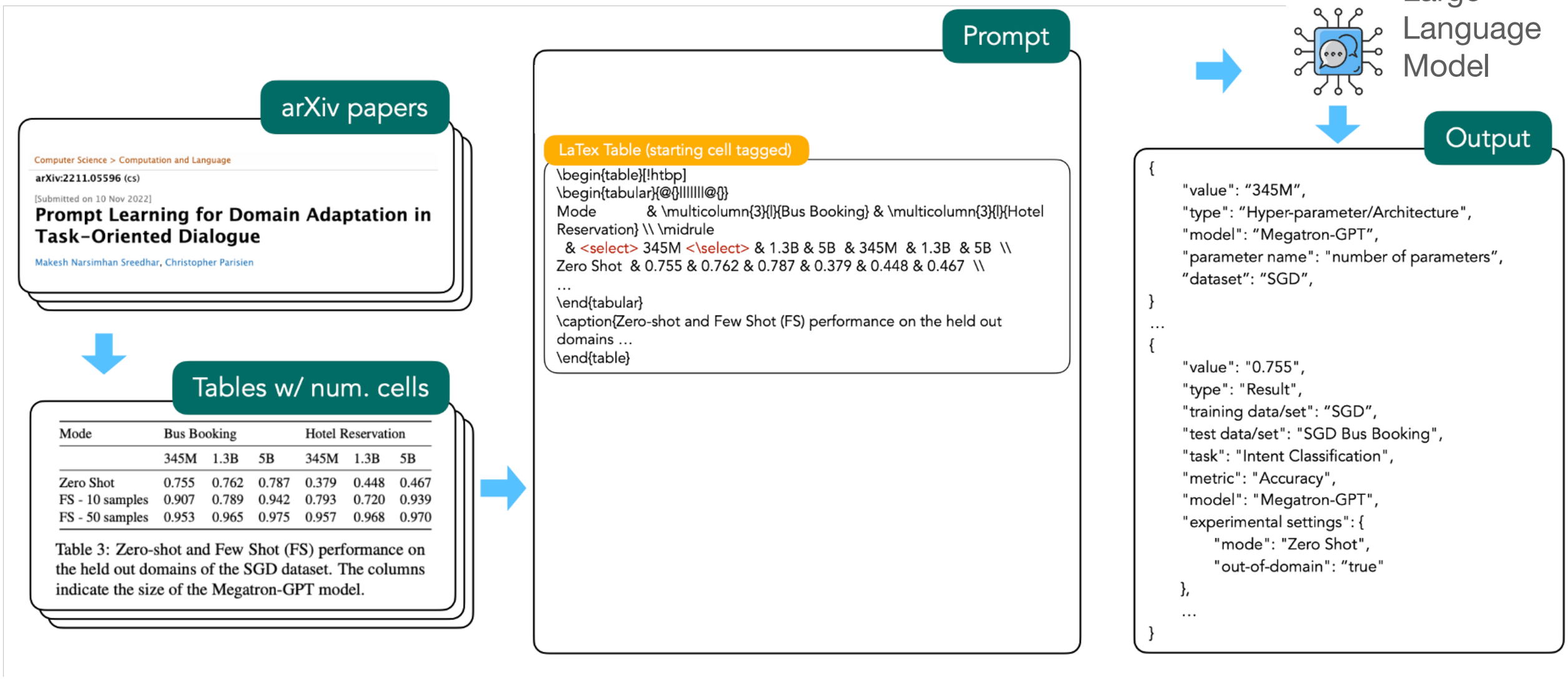
	MLTAB. (ours)	CHEMTAB. (ours)	DISCOMAT (Gupta et. al. 2023)	SWDE (Hao et. al. 2011)
# cell types	4	6	2	8
# papers (web.)	25	16	2,536	80
# tables (pages)	122	26	5,883	1,600
# anno. records	3,792	1,498	58,481	1,600

Benchmarking Schema-Driven IE



	MLTAB. (ours)	CHEMTAB. (ours)	DISCOMAT (Gupta et. al. 2023)	SWDE (Hao et. al. 2011)
# cell types	4	6	2	8
# papers (web.)	25	16	2,536	80
# tables (pages)	122	26	5,883	1,600
# anno. records	3,792	1,498	58,481	1,600

Schema-Driven IE



Schema-Driven IE

arXiv papers

Computer Science > Computation and Language

arXiv:2211.05596 (cs)

[Submitted on 10 Nov 2022]

Prompt Learning for Domain Adaptation in Task-Oriented Dialogue

Makesh Narsimhan Sreedhar, Christopher Parisien

Tables w/ num. cells

Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

Table 3: Zero-shot and Few Shot (FS) performance on the held out domains of the SGD dataset. The columns indicate the size of the Megatron-GPT model.

Prompt

Retrieved paragraphs

In this work, we explore the task of intent classification using these large language models and ptuning. Generative methods …

LaTex Table (starting cell tagged)

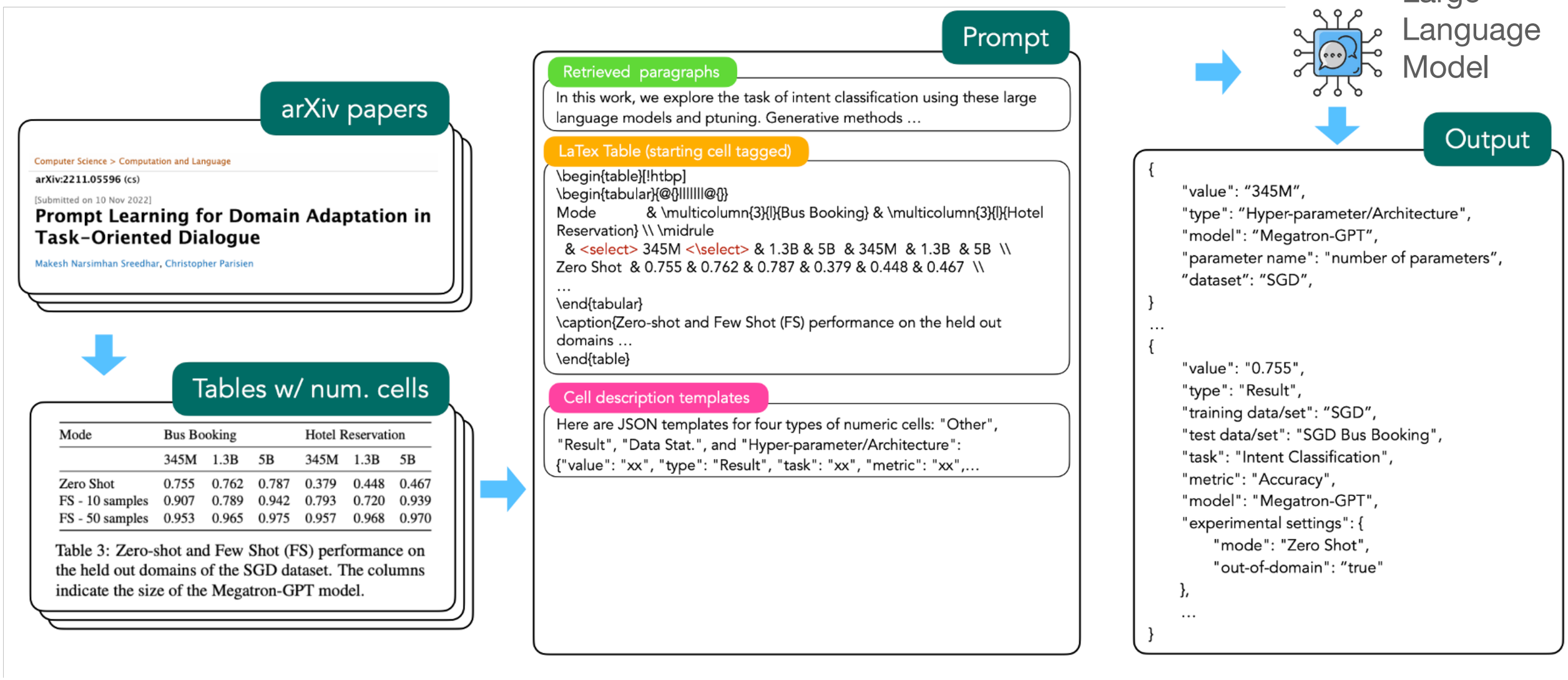
```
\begin{table}[!htbp]
\begin{tabular}{@{}lllllll@{}}
Mode          & \multicolumn{3}{l}{Bus Booking} & \multicolumn{3}{l}{Hotel
Reservation} \\ \midrule
  & <select> 345M <\select> & 1.3B & 5B  & 345M  & 1.3B  & 5B  \\
Zero Shot  & 0.755 & 0.762 & 0.787 & 0.379 & 0.448 & 0.467  \\
…
\end{tabular}
\caption{Zero-shot and Few Shot (FS) performance on the held out
domains …
\end{table}
```

Large Language Model

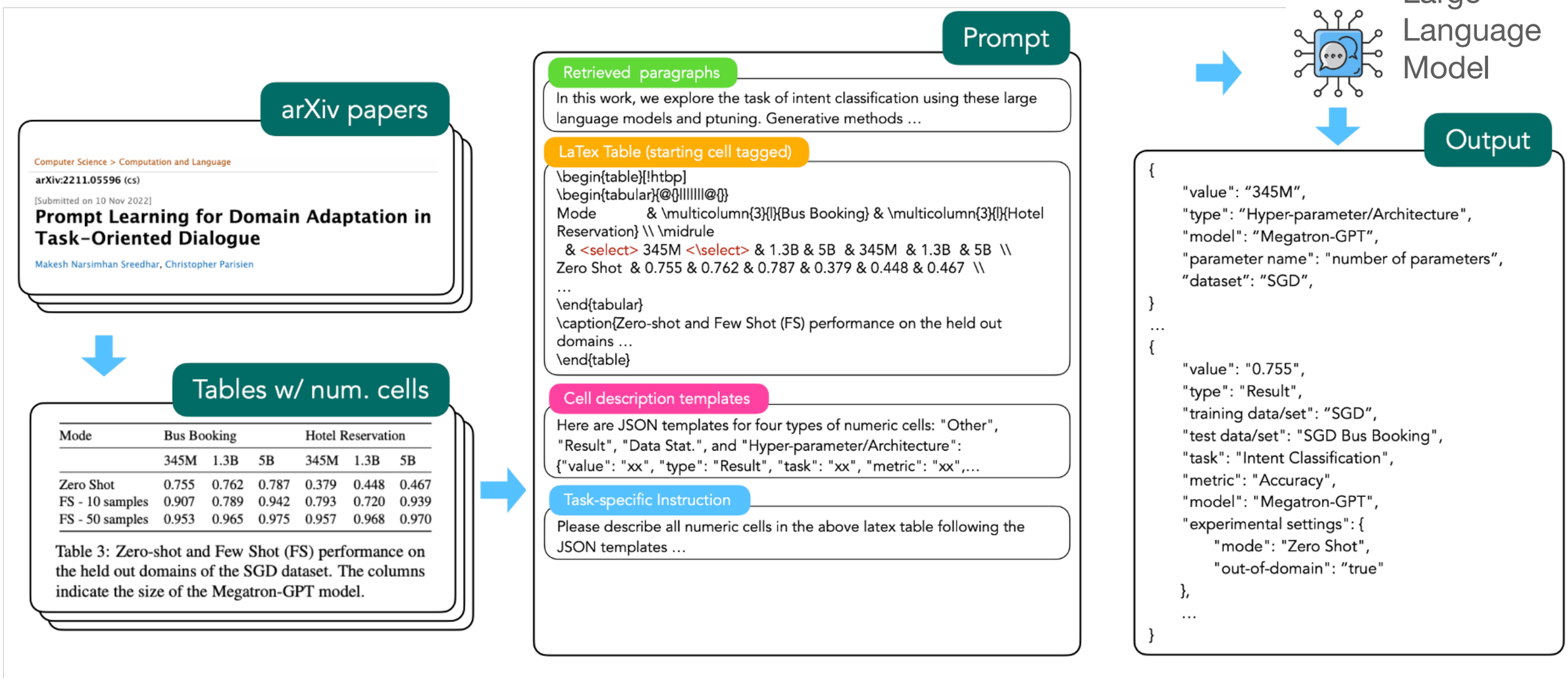
Output

```
{
    "value": "345M",
    "type": "Hyper-parameter/Architecture",
    "model": "Megatron-GPT",
    "parameter name": "number of parameters",
    "dataset": "SGD",
}
…
{
    "value": "0.755",
    "type": "Result",
    "training data/set": "SGD",
    "test data/set": "SGD Bus Booking",
    "task": "Intent Classification",
    "metric": "Accuracy",
    "model": "Megatron-GPT",
    "experimental settings": {
        "mode": "Zero Shot",
        "out-of-domain": "true"
    },
    …
}
```

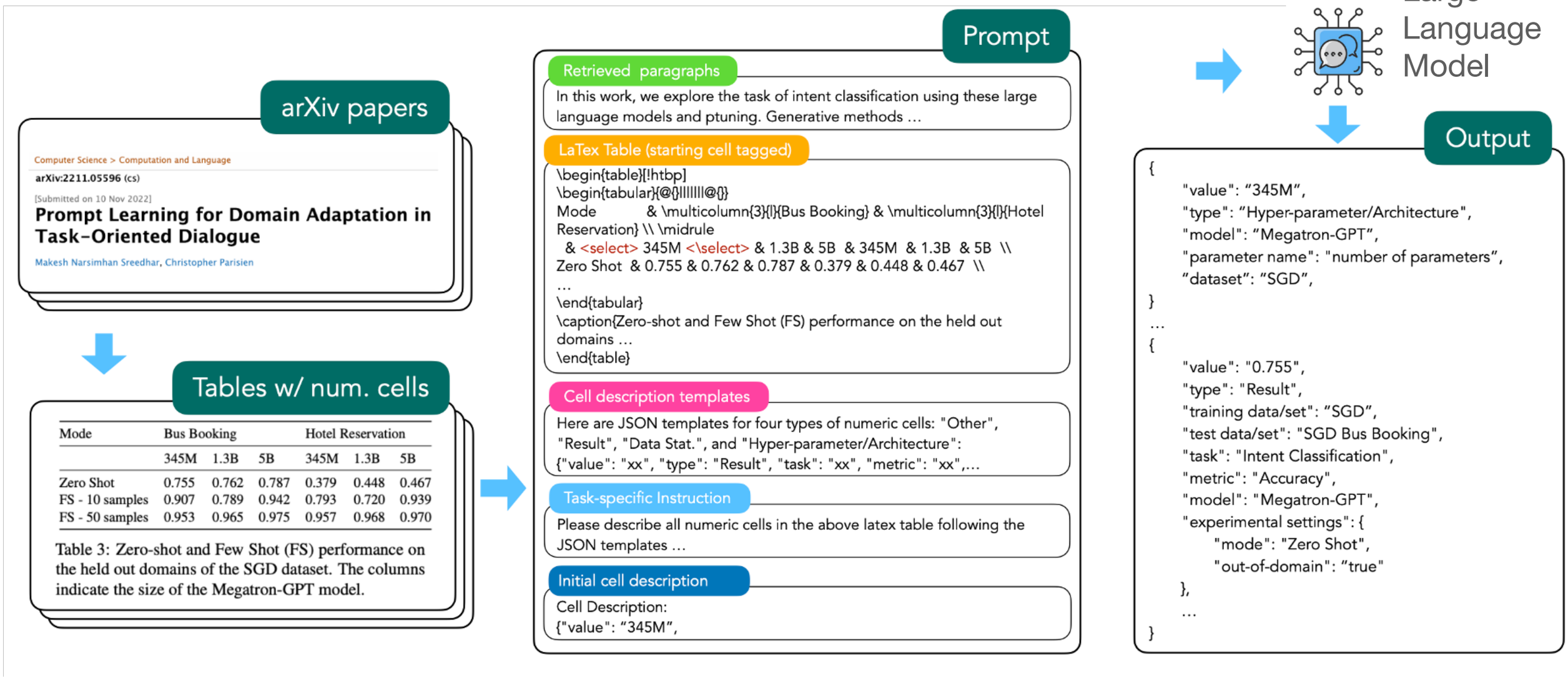
Schema-Driven IE



Schema-Driven IE



Schema-Driven IE



Error Recovery

- **Challenge:** Language models visit table cells in disorganized manner.

1. Raw Record Output

Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

Table 3: Zero-shot and Few Shot (FS) performance on the held out domains of the SGD dataset. The columns indicate the size of the Megatron-GPT model.

{"value": "345M", "type": "Hyper-params.", "model": "Mega..."}
{"value": "1.3B", "type": "Hyper-params.", "model": "Mega..."}
{"value": "5B", "type": "Hyper-params.", "model": "Mega..."}
{"value": "345M", "type": "Hyper-params.", "model": "Mega..."}
...
{"value": "0.755", "type": "Result", "training data": "SGD", ...}
{"value": "0.907", "type": "Result", "training data": "SGD", ...}
{"value": "0.953", "type": "Result", "training data": "SGD", ...}

Error Recovery

- **Challenge:** Language models visit table cells in disorganized manner.

2. Record Order Checking

Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

Table 3: Zero-shot and Few Shot (FS) performance on the held out domains of the SGD dataset. The columns indicate the size of the Megatron-GPT model.

Follows the instructed "left-right, top-down" order

{"value": "345M", "type": "Hyper-params.", "model": "Mega..."}
{"value": "1.3B", "type": "Hyper-params.", "model": "Mega..."}
{"value": "5B", "type": "Hyper-params.", "model": "Mega..."}
{"value": "345M", "type": "Hyper-params.", "model": "Mega..."}
...
{"value": "0.755", "type": "Result", "training data": "SGD", ...}
{"value": "0.907", "type": "Result", "training data": "SGD", ...}
{"value": "0.953", "type": "Result", "training data": "SGD", ...}

Does not follow the instructed order (truncated)

Error Recovery

- **Challenge:** Language models visit table cells in disorganized manner.

3. Record Error Recovery

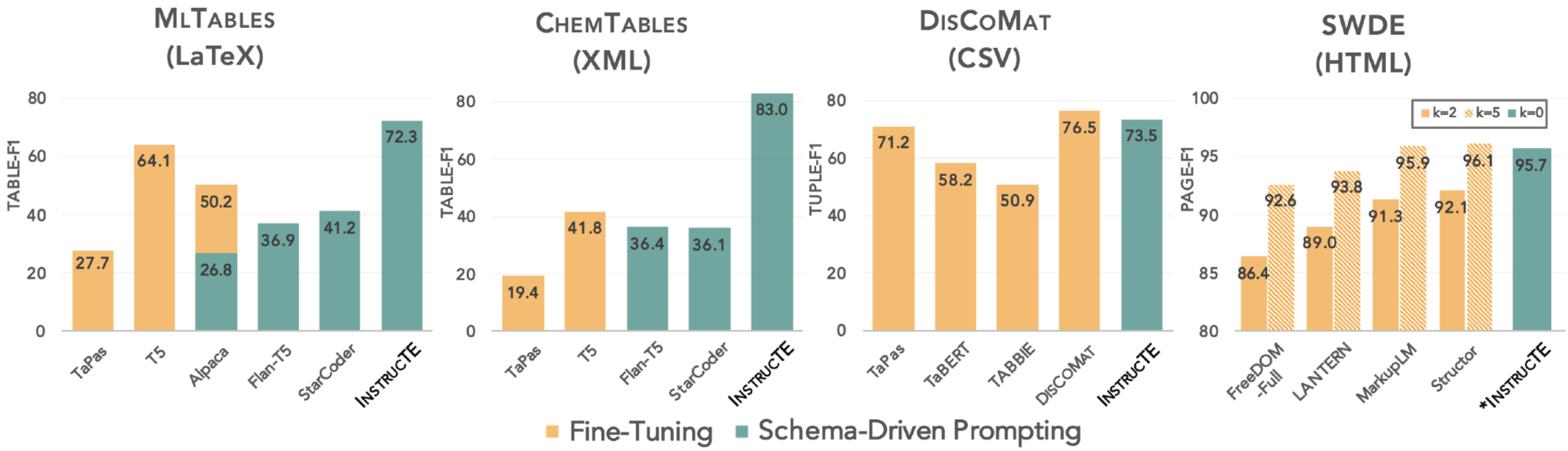
Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

Table 3: Zero-shot and Few Shot (FS) performance on the held out domains of the SGD dataset. The columns indicate the

{"value": "345M", "type": "Hyper-params.", "model": "Mega..."}
{"value": "1.3B", "type": "Hyper-params.", "model": "Mega..."}
{"value": "5B", "type": "Hyper-params.", "model": "Mega..."}
{"value": "345M", "type": "Hyper-params.", "model": "Mega..."}
...
{"value": "0.755", "type": "Result", "training data": "SGD", ...}
{"value": "0.762", "type":

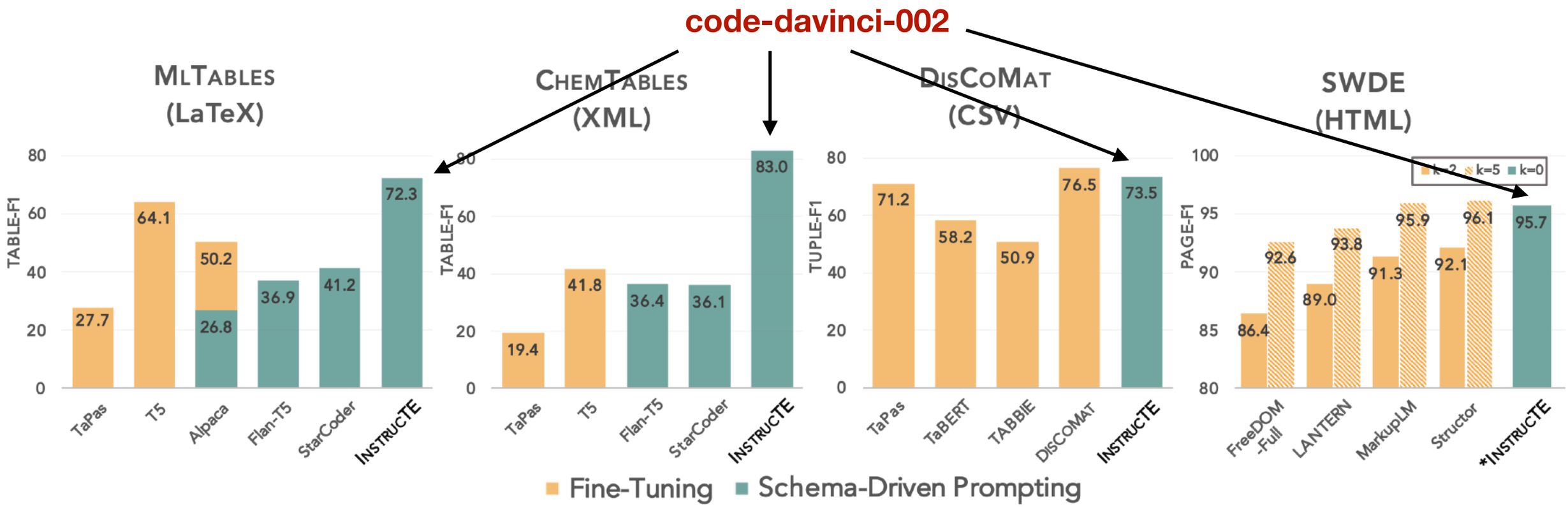
Append the next cell (following the instructed order) and re-prompt the model

Results



- Competitive with SOTA supervised models
  - No Labels or Custom Pipelines
  - Works on Diverse Data Formats and Domains
  - Only Human Supervision is Extraction Schema

Results



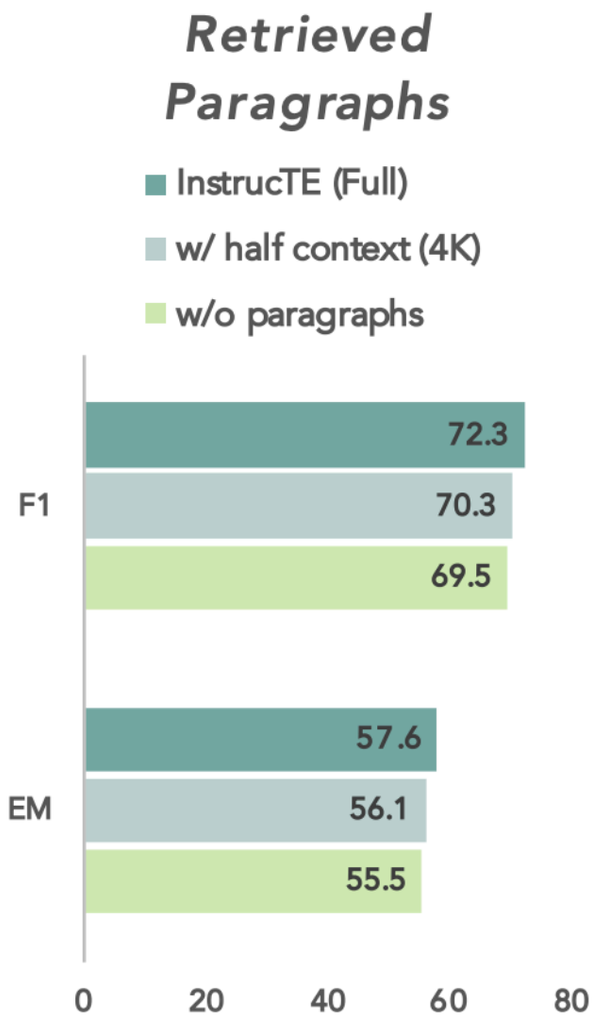
- Competitive with SOTA supervised models
  - No Labels or Custom Pipelines
  - Works on Diverse Data Formats and Domains
  - Only Human Supervision is Extraction Schema

Distilling Table Extractors

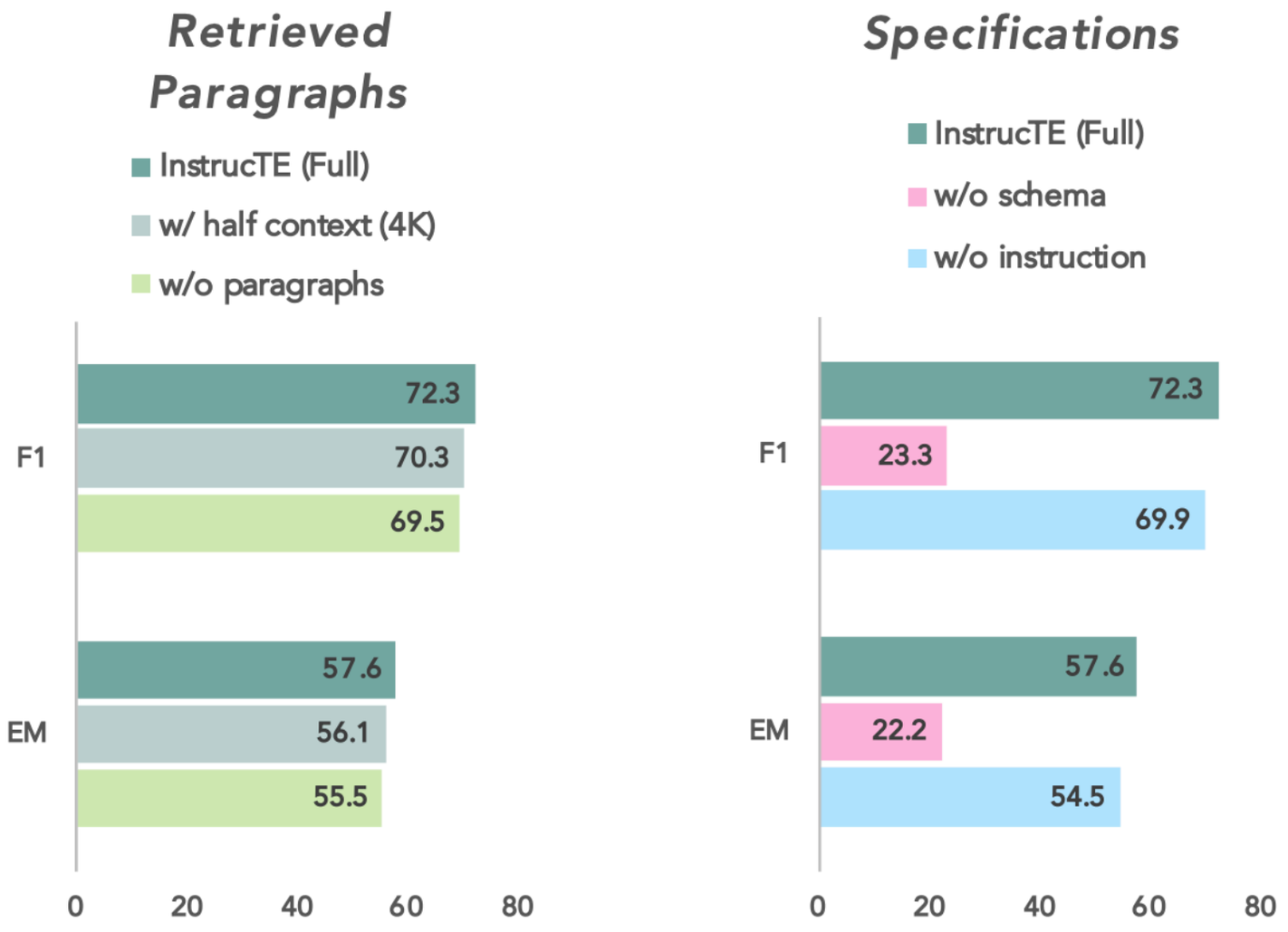
	Model (GPU hours)	Token-Level F_1			EM		
		P	R	F_1	P	R	F_1
Teacher	code-davinci-002	74.1	71.8	72.3	59.4	56.9	57.6
Student	LLaMA-7B (50h)	74.1	67.6	69.1	56.8	53.4	54.3
	Alpaca-7B (50h)	72.7	64.8	67.5	56.1	50.0	52.0
	T5-11B (380h)	75.8	71.4	73.2	60.3	56.7	58.1

- Not difficult to distill student table extraction models with performance as good as the teacher.

Ablation Study



Ablation Study

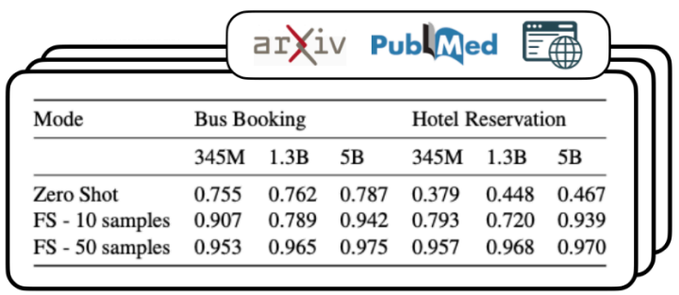


- Schemas are crucial

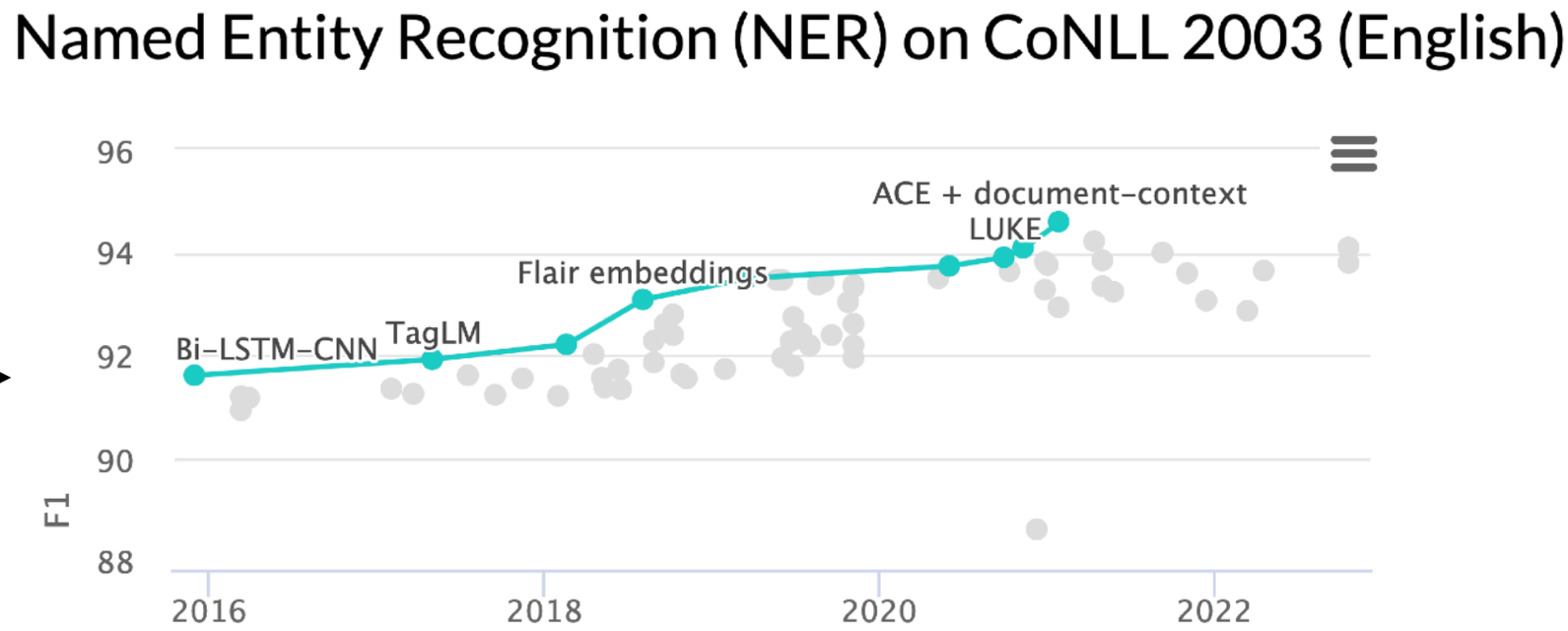
Ablation Study

- Schemas are crucial
- Error recovery has strong performance while minimizing API costs.

Downstream Task: ML Leaderboards (Kadras et. al. 2020)

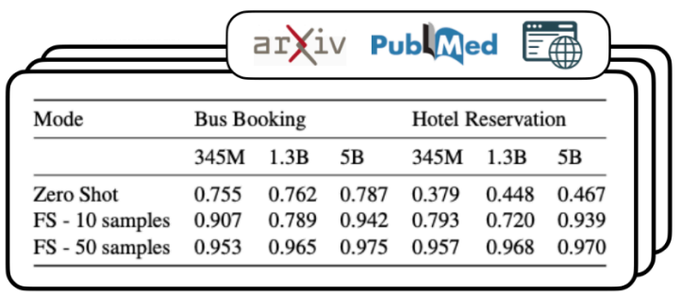


Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

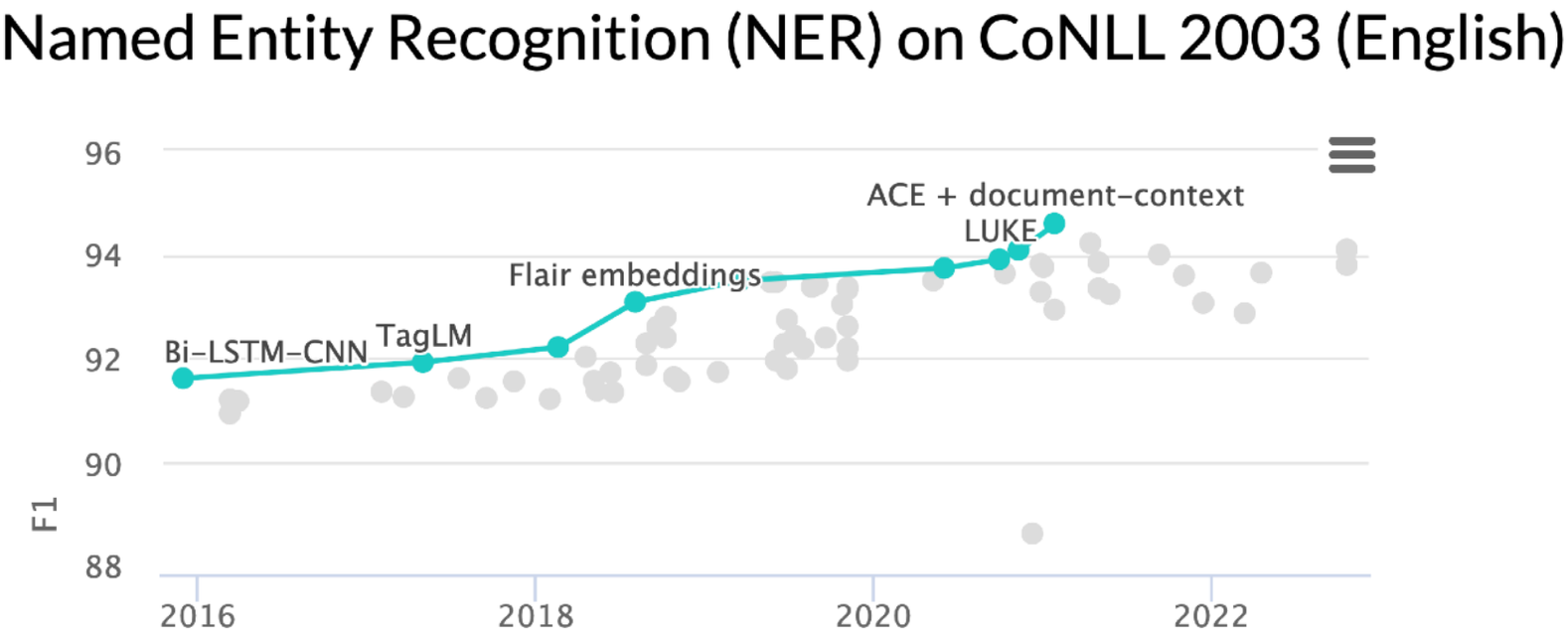


- Requires Additional Linking Steps

Downstream Task: ML Leaderboards (Kadras et. al. 2020)



Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970



- Requires Additional Linking Steps
  - **Performance comparable to custom Pipeline w/ labeled examples**

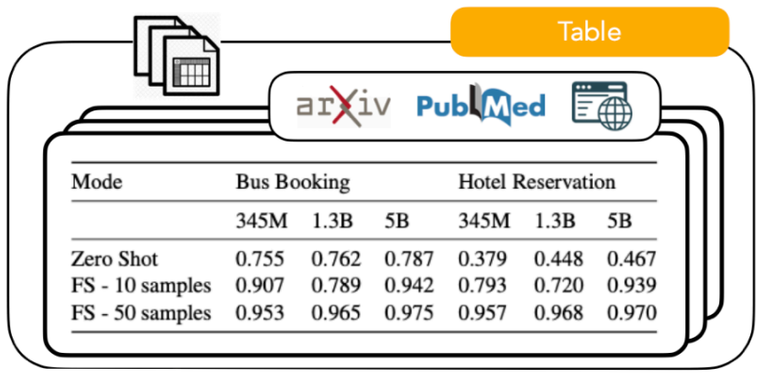
Method	Micro-Average		
	P	R	F_1
AXCELL	**25.4**	18.4	21.3
INSTRUCTE	20.1	20.8	20.5
INSTRUCTE+	23.9	**21.2**	**22.4**

Takeaways

(1) Schema-Driven Information Extraction

Schema-Driven Information Extraction from Heterogeneous Tables
Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Alan Ritter
arXiv preprint (2023)

- Key idea: Only supervision is **extraction schema**
- Strong performance across multiple domains



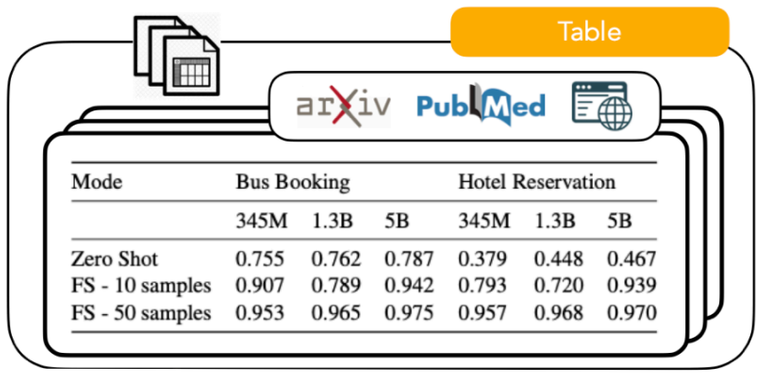
Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

Takeaways

(1) Schema-Driven Information Extraction

Schema-Driven Information Extraction from Heterogeneous Tables
Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Alan Ritter
arXiv preprint (2023)

- Key idea: Only supervision is **extraction schema**
- Strong performance across multiple domains

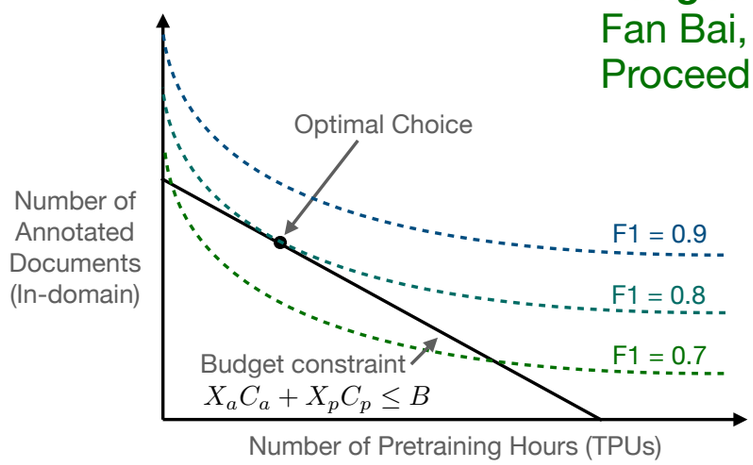


Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

(2) Computation vs. Annotation

Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models
Junmo Kang, Wei Xu and Alan Ritter
To Appear at ACL 2023

Pre-train or Annotate? Domain Adaptation with a Constrained Budget
Fan Bai, Alan Ritter and Wei Xu
Proceedings of EMNLP 2021

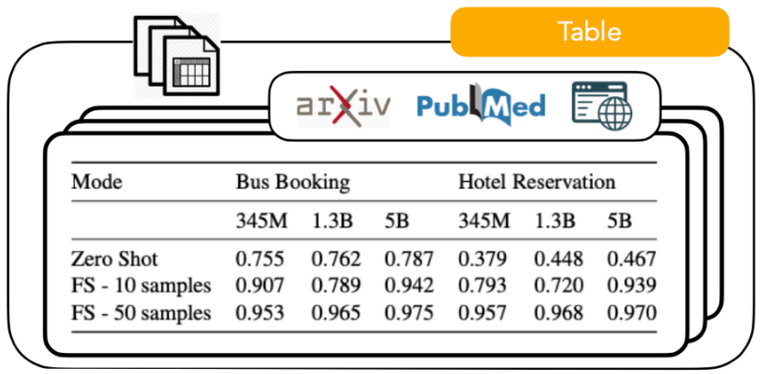


Takeaways

(1) Schema-Driven Information Extraction

Schema-Driven Information Extraction from Heterogeneous Tables
Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Alan Ritter
arXiv preprint (2023)

- Key idea: Only supervision is **extraction schema**
- Strong performance across multiple domains

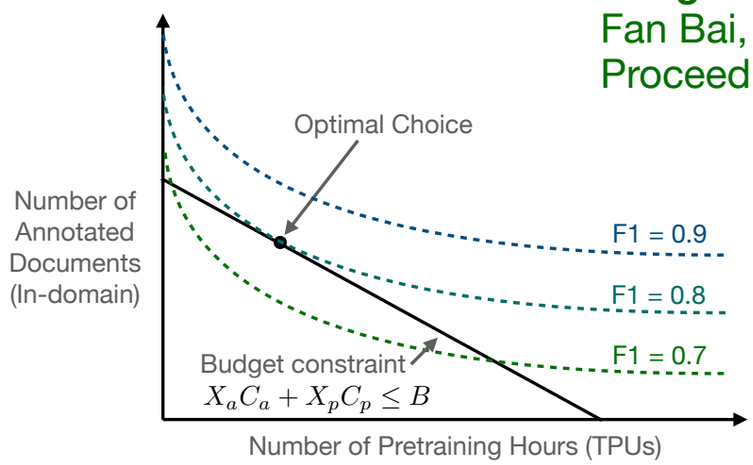


Mode	Bus Booking			Hotel Reservation		
	345M	1.3B	5B	345M	1.3B	5B
Zero Shot	0.755	0.762	0.787	0.379	0.448	0.467
FS - 10 samples	0.907	0.789	0.942	0.793	0.720	0.939
FS - 50 samples	0.953	0.965	0.975	0.957	0.968	0.970

(2) Computation vs. Annotation

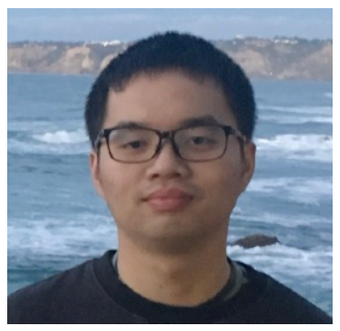
Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models
Junmo Kang, Wei Xu and Alan Ritter
To Appear at ACL 2023

Pre-train or Annotate? Domain Adaptation with a Constrained Budget
Fan Bai, Alan Ritter and Wei Xu
Proceedings of EMNLP 2021



Thank you!

Sponsors:
NSF DARPA


Fan Bai


Junmo Kang