Directions Towards Efficient and Automated Data Wrangling with Large Language Models

Zeyu Zhang, Paul Groth, Iacer Calixto, Sebastian Schelter

13 May 2024

University of Amsterdam

z.zhang2@uva.nl

Motivation: Data Wrangling with Large Language Models (LLMs)

- Huge **potential of LLMs for long-standing data wrangling tasks** such as entity matching, missing value imputation and error detection [1, 2]

- **Automation and scalability challenges** (e.g. for data wrangling services in the cloud)
    - Manual few-prompt selection from [1] **not automatable and scalable**
    - **Disadvantages of automatable alternatives** such as fully fine-tuning a model per customer
        - **High storage costs** (for copies of model parameters)
        - **High computational costs** (for model training)

→ We need **parameter- and compute-efficient** ways to employ LLMs for data wrangling

[1] Narayan et al.: Can Foundation Models Wrangle Your Data?, VLDB'22
[2] Fernandez et al.: How large language models will disrupt data management, VLDB'23

Our Contributions

- **Extended study on parameter-efficient finetuning (PEFT) of LLMs** for data wrangling
  - Four popular PEFT methods, three baselines, three LLM variants, ten benchmark datasets
  - Measure training and inference time in addition to prediction quality

- **Vision for zero-shot entity matching**
  - Exploration of a zero-shot setting for entity matching to further reduce deployment costs

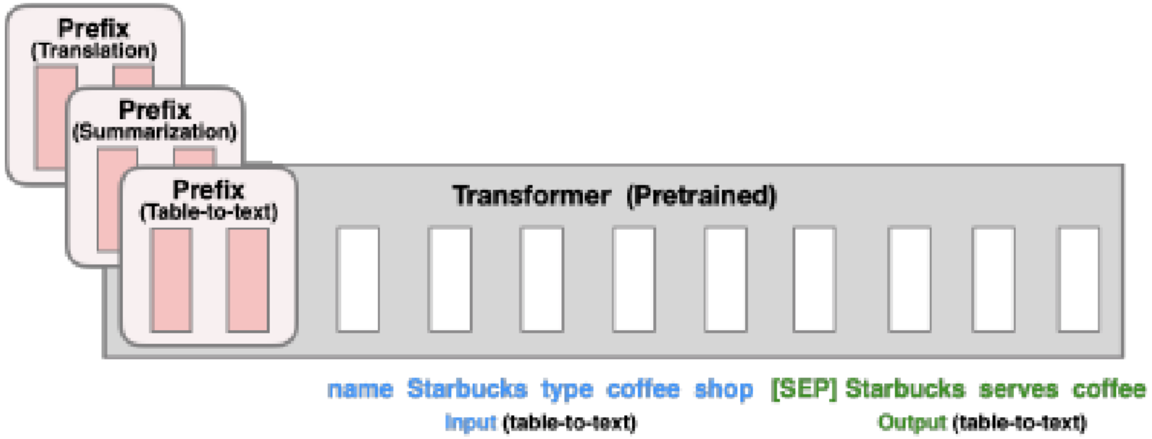
- **Reproducible benchmark**
  - Code and experimental results available at https://github.com/Jantory/cpwrangle

Study on parameter-efficient fine-tuning of LLMs for Data Wrangling

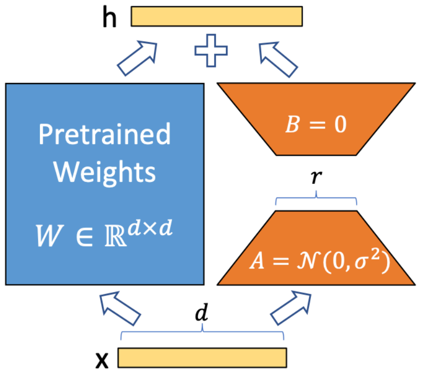
Parameter Efficient Fine-Tuning

Transfer Learning techniques for LLMs

- **Manual prompt engineering** -- no training (+), hard to automate (-)
- **Full finetuning (FT)** -- high performance (+), requires substantial computational resources (-)
- **Parameter Efficient Tuning (PEFT)** -- fewer parameters trained (+), on par performance (+)



Prefix-tuning[1]



LoRA adapter[2]

[1] Li et al., “Prefix-tuning: Optimizing continuous prompts for generation,” ACL’21.
[2] Huetal.,“LoRA:Low-RankAdaptationofLargeLanguageModels,” ICLR’22.

Results on Prediction Quality

How does prediction quality vary among different PEFT methods and base models?

LLM	Method	# of Parameter Updates	Mean Predictive Score
GPT3 (175B)	Zero-Shot	-	66.71
-	AutoML	-	76.88
T5-small (60.5M)	Prompt	48K	81.94
	P-tune	212K	80.11
	Prefix	309K	67.66
	LoRA	296K	90.96
	Finetune	60,500K	89.95
T5-base (223M)	Prompt	67K	81.22
	P-tune	312K	85.09
	Prefix	914K	84.49
	LoRA	892K	92.03
	Finetune	223,000K	90.36
T5-large (783M)	Prompt	74K	82.04
	P-tune	369K	76.62
	Prefix	2,435K	88.65
	LoRA	2,362K	**92.24**
	Finetune	770,000K	Train Failed

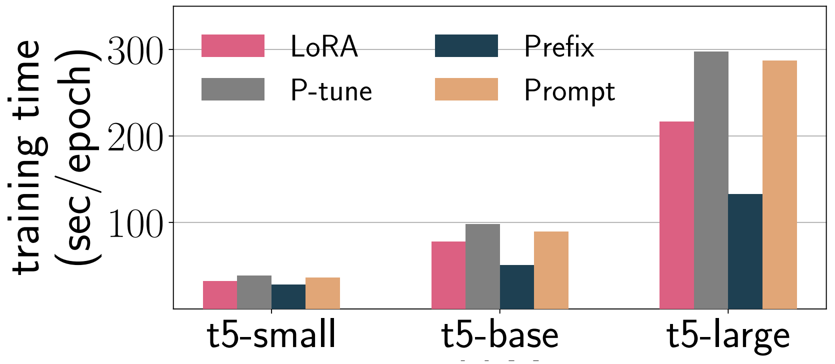
Evaluated **four PEFT methods** (Prompt, P-tune, Prefix, LoRA) on **three variants of Google's T5 model** on benchmark data from Narayan et al.

Findings:

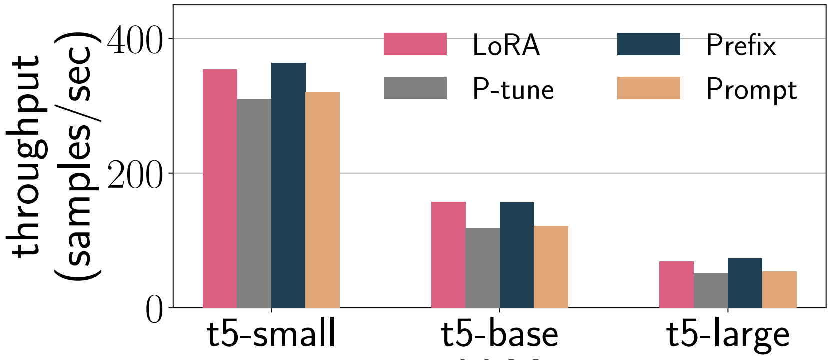
- PEFT methods outperform GPT3 baseline and AutoML in many settings
- LoRA provides highest performance
- Applying PEFT methods to larger models provides higher performance

Results on Computational Efficiency

How does computational efficiency vary among different PEFT methods and base models?



Training time per epoch on AMGO dataset



Mean **inference throughput** over all datasets

Training Times for FT on AMGO Dataset: 38s, 109s, and 312s, respectively.

Findings:

- Only minor differences in training and inference times between PEFT methods, parameter size has highest impact
- **PEFT methods designed for parameter efficiency** (two orders of magnitude less parameters than full finetuning), **but not for compute efficiency!**

Even the fastest method Prefix-Tuning is only twice as fast as full fine-tuning on t5-base

Vision for Zero-Shot Entity Matching

Towards Zero-Shot Entity Matching

- **PEFT methods highly automatable** and very **parameter-efficient**

- PEFT methods still incur **high training costs**
- Potential show-stopper for use cases that require many custom models (e.g., a model per customer in a cloud service)

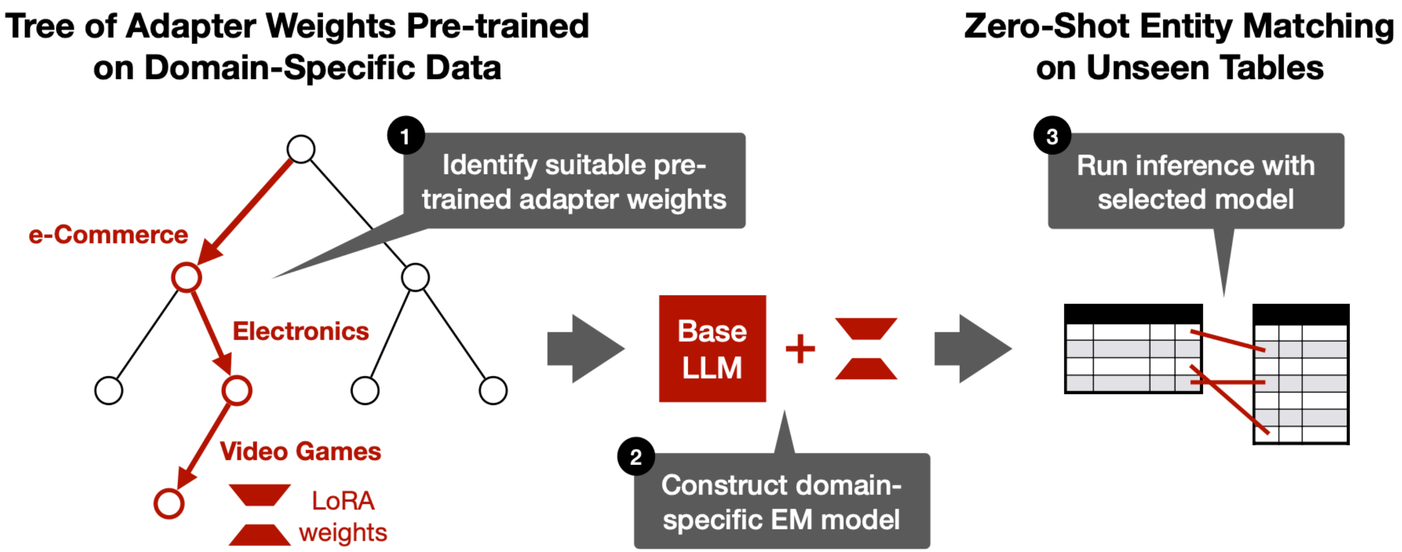
- Desideratum: **Models that can be applied without fine-tuning in a zero-shot manner**

- Found evidence for **zero-shot potential of LLMs for Entity Matching!**

Target dataset	**GPT-3 (175B)**	**T5-base (223M)**			
		LoRA	pretrained on	Prompt	pretrained on
iTunes-Amazon	65.90	**94.73**	Beer	91.52	Walmart-Amazon
Beer	78.60	**93.33**	DBLP-Google	87.50	Walmart-Amazon
Fodors-Zagats	87.50	**100.00**	iTunes-Amazon	97.67	Walmart-Amazon
Walmart-Amazon	60.60	**62.92**	Beer	45.51	DBLP-Google
Amazon-Google	54.30	**62.75**	DBLP-Google	61.85	Walmart-Amazon
DBLP-ACM	93.50	93.73	DBLP-Google	**96.25**	DBLP-Google
DBLP-Google	64.60	**88.96**	DBLP-ACM	81.34	Walmart-Amazon

Vision for Zero-Match

- **A hierarchical soft prompt/adapter tree for the zero-shot setting**



Thanks! Questions?

- **Summary**
  - Compared prediction quality PEFT methods for data wrangling.
  - Compared computational efficiency of PEFT methods for data wrangling.
  - Vision for zero-shot entity matching based on the PEFT approach.

- **Checkout our paper for more details!**
- Code and experimental results available at https://github.com/Jantory/cpwrangle

- Contact me at z.zhang2@uva.nl