In-Context-Learning For Few-Shot Adaptation of LLMs

Sunita Sarawagi
IIT Bombay
sunita@iitb.ac.in

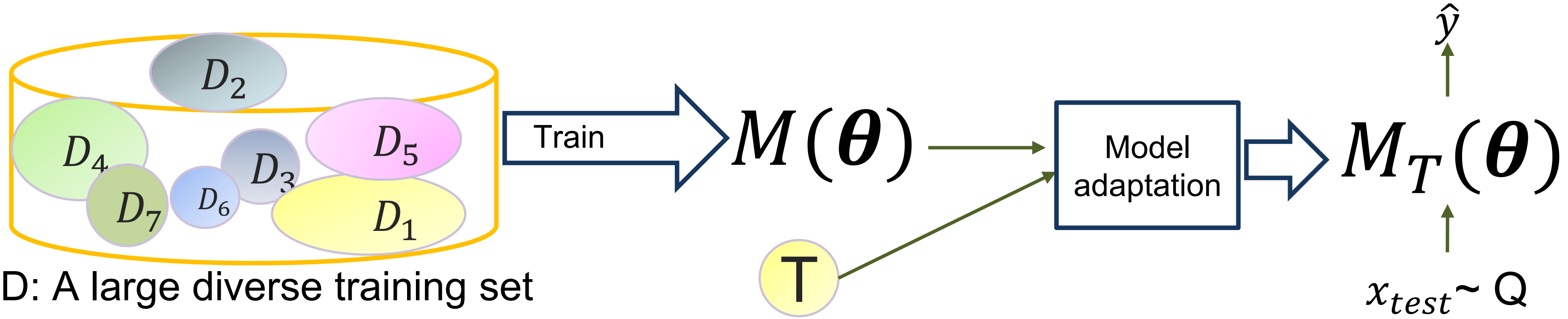
Few-shot Model Adaptation

Given a trained model $\mathrm{M}(\boldsymbol{\Theta})$ on a large diverse dataset D

User interested in a target task $\mathbb{Q}$, gives a **small** labelled data

T: $\{x_1 \quad y_1 \quad x_2 \quad y_2 \quad x_3 \quad y_3 \quad \dots\dots \quad x_k \quad y_k\} \sim$ Q

Model adaptation = Find the best blend of $\mathrm{M}(\boldsymbol{\Theta})$ and T



D: A large diverse training set

D may contain a part D_j that is relevant to target task, but it is not explicitly identified.

Example: model adaptation

$M(\boldsymbol{\theta})$	**Target task**
Conversation assistant	→ New users
Text-to-SQL system	→ Private pharma database
A large language model	→ NER on physics documents

Conversation assistant

→ New users

Text-to-SQL system

Concerts

Football

Travel

Countries

Books

→ Private pharma database

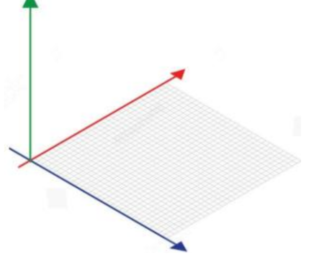
A large language model

WIKIPEDIA

Diverse topics and tasks

→ NER on physics documents

The challenges of model adaptation



- Accuracy: requires delicate balance.

- Robustness: Small T leads to high variance across T~Q
  - Cannot introduce additional instabilities, e.g. due to ordering
- Efficiency: $\boldsymbol{\Theta}$ is large
  - Time of adaptation: Offline Vs Online (On-the-fly)
  - Test time: time for using after adaptation.

Model adaptation over the Ages

Thousands of paper under variants like domain adaptation, transfer-learning, few-shot learning..

Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lo
Ciprian Chelba · A. Acero · Computer Science · Computer Speech and Language · 2006

Domain adaptation of natural language processing systems
Fernando C Pereira · John Blitzer · Computer Science, Linguistics · 2008

Domain Adaptation of Conditional Probability Models Via Feature Subsetting
Sandeepkumar Satpal · Sunita Sarawagi · Computer Science · European Conference on Principles of Data Mining... · 17 September 2007

Domain adaptation of information extraction models
Rahul Gupta · Sunita Sarawagi · Computer Science · SGMD · 20 March 2009

Structured Case-based Reasoning for Inference-time Adaptation of Text-to-SQL parsers
Abhijeet Awasthi · Soumen Chakrabarti · Sunita Sarawagi · Computer Science · AAAI Conference on Artificial Intelligence · 10 January 2023

Adversarial Discriminative Domain Adaptation
Eric Tzeng · Judy Hoffman · Kate Saenko · Trevor Darrell · Computer Vision and Pattern Recognition · 17 February 2017

LoRA: Low-Rank Adaptation of Large Language Models
J. E. Hu · Yelong Shen · +4 authors · Weizhu Chen · Computer Science · International Conference on Learning... · 17 June 2021

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks
Chelsea Finn · P. Abbeel · S. Levine · Computer Science · International Conference on Machine Learning ·

Unsupervised Domain Adaptation by Backpropagation
Yaroslav Ganin · V. Lempitsky · Computer Science · International Conference on Machine Learning · 26 September 2014

Deep Transfer Learning with Joint Adaptation Networks
Mingsheng Long · Hanhua Zhu · Jianmin Wang · Michael I. Jordan · International Conference on Machine Learning · 21 May 2016

Model adaptation methods

- Fine-tuning and its variants
- Mixture of experts
- Task-vectors
- Matching-based methods

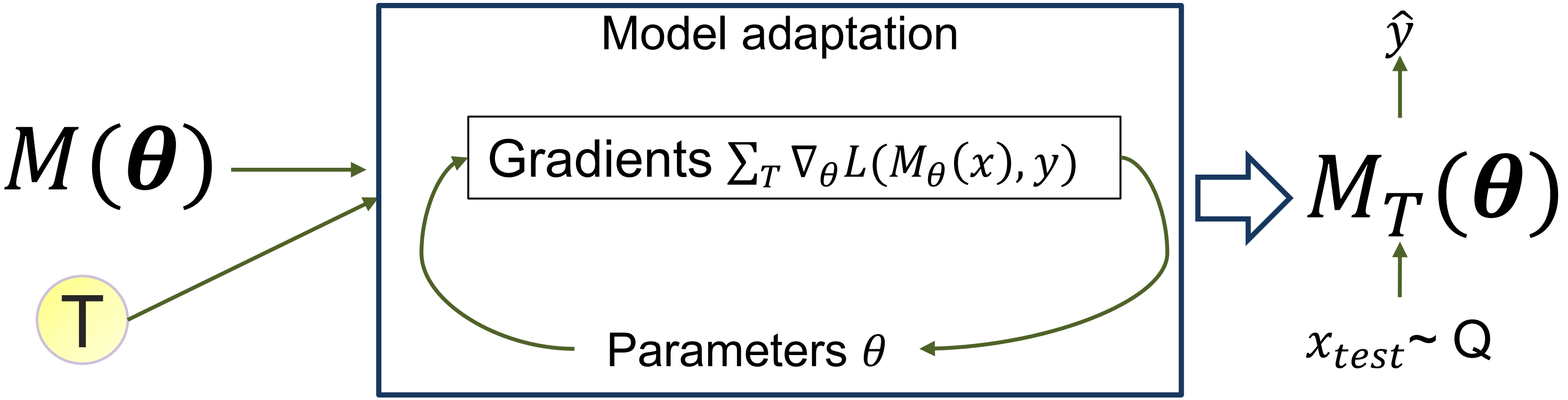
Each category explicitly planned for adaptation either when

- Training $\mathrm{M}(\boldsymbol{\Theta})$, or when
- Adapting with T

Fine-tune parameters via gradient descent on T

Training of $M(\theta)$:
Normal loss on D

Adaptation with T:
Update θ via gradient descent on T



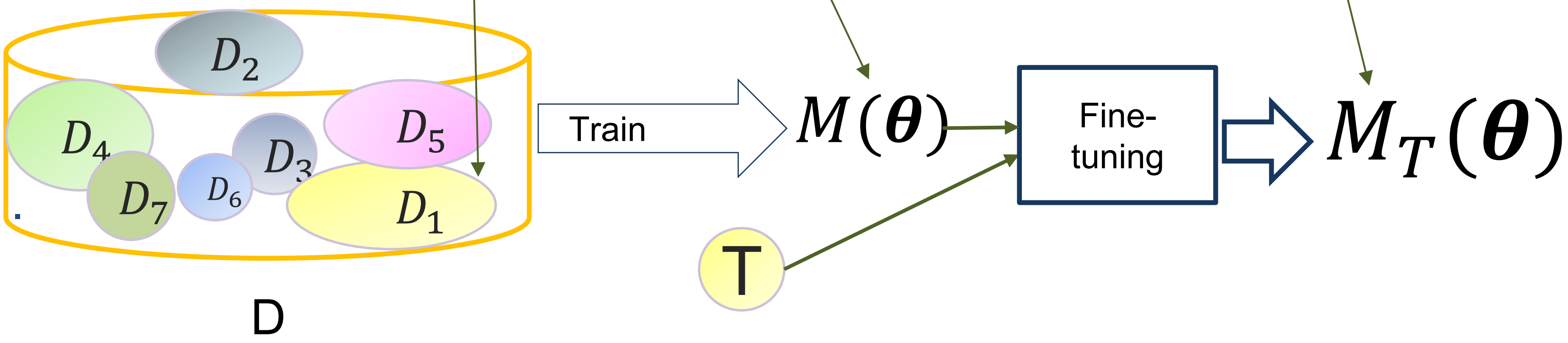
Explicit update of model parameters Θ to minimize loss on target data T
Generally provides good accuracy for modest sized T

Does fine-tuning define the accuracy ceiling?

D is diverse, and of mixed relevance to target

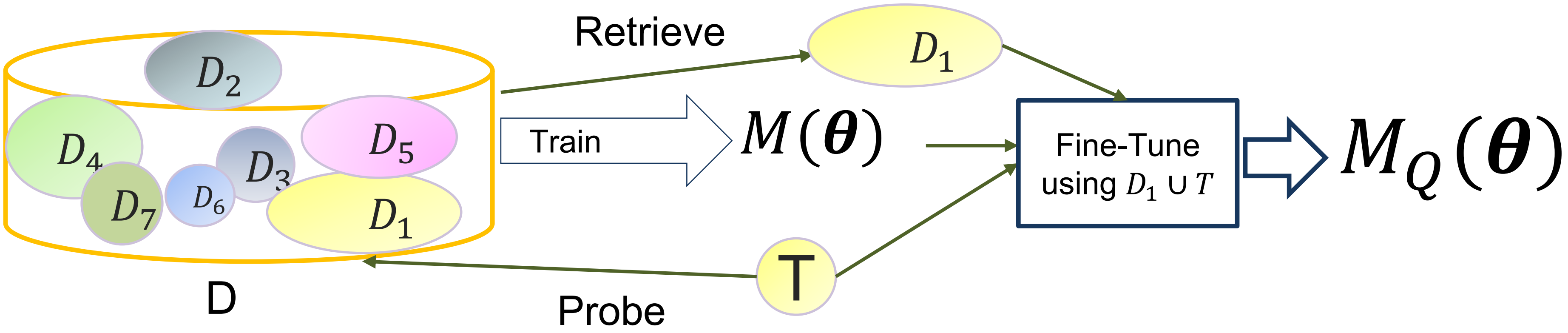
$M(\boldsymbol{\theta})$ trained on D, may not preserve subparts

such that fine-tuning with T recalls relevant parts

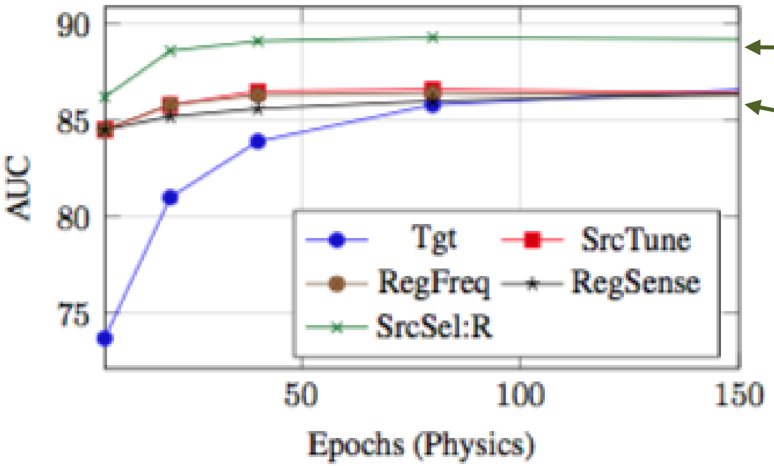


An accuracy ceiling: Revisit the source D

- Identify relevant D_1 explicitly from D, Retrieval problem.
- Fine-tune with $D_1 \cup \mathrm{T}$.



Revisit the source for adapting word embeddings



Revisit the source

Fine-tuning variants

Revisiting source significantly better than best fine-tuning.

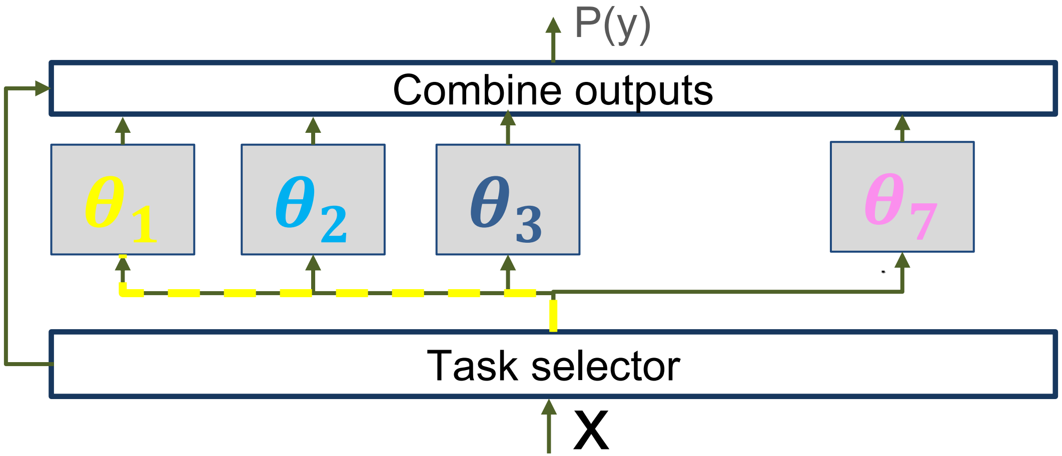
Not practical. Cannot revisit source

Useful intuition: important to maintain task identities for later specialization

Topic-Sensitive Attention on Generic Corpora Corrects Sense Bias in Pretrained Embeddings. Piratla, Sarawagi and Chakrabarti In *ACL*, 2019.

A practical approximation: mixture of experts

Model: mixture of experts with a task selector



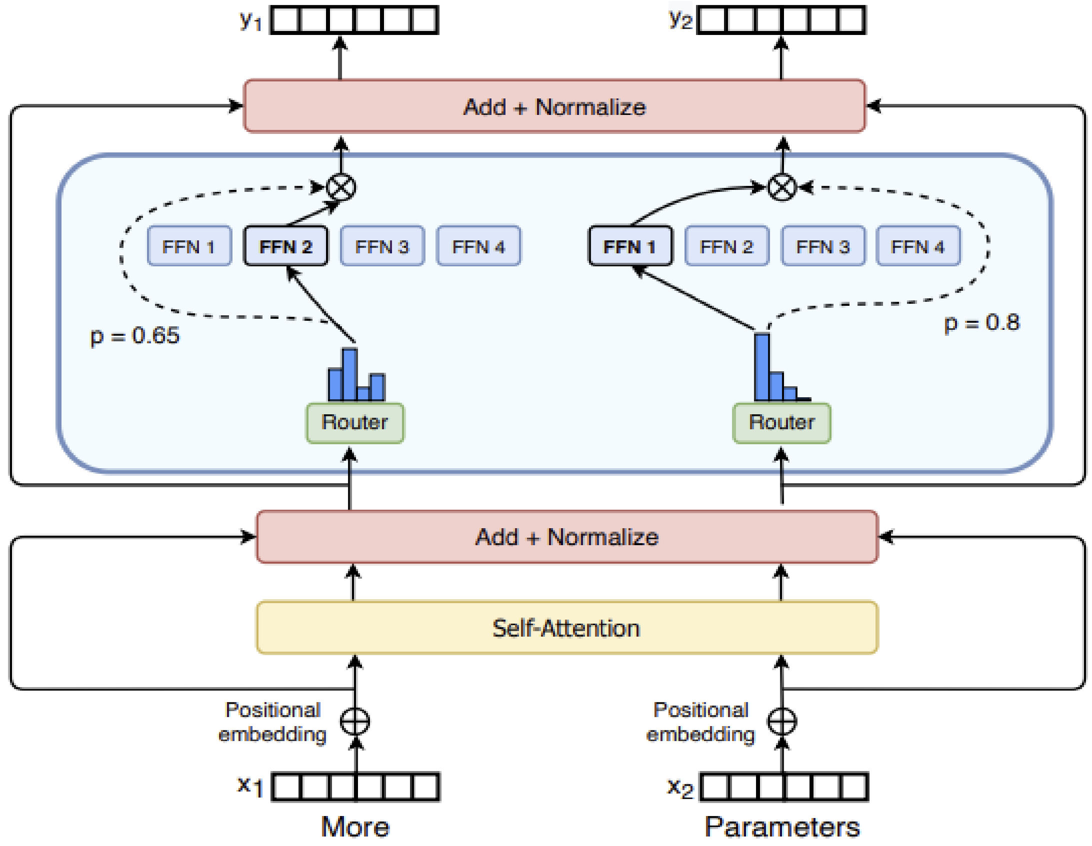
Training of $M(\theta)$: Task selector supervised by tasks in D

Adaptation with T: recognize the most relevant expert

Separate parameters for each task may be too much.

Multi-Source Domain Adaptation with Mixture of Experts. EMNLP 2018. Guo, Shah, Barzilay

Modern Mixture of Experts: Switch Transformers

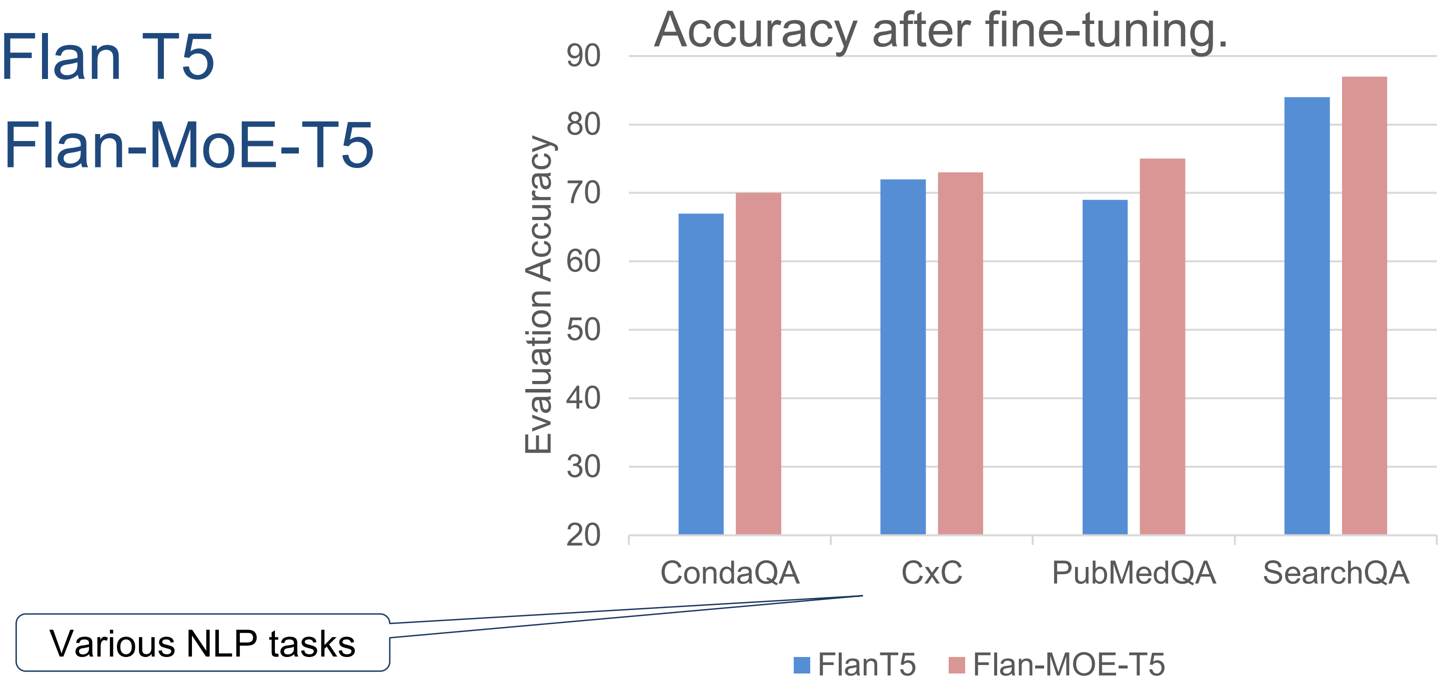


Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. JMLR 2022 Fedus, Zoph, Shazeer

Greater accuracy gains on fine-tuning MoEs

Base model: Flan T5

MoE variant: Flan-MoE-T5

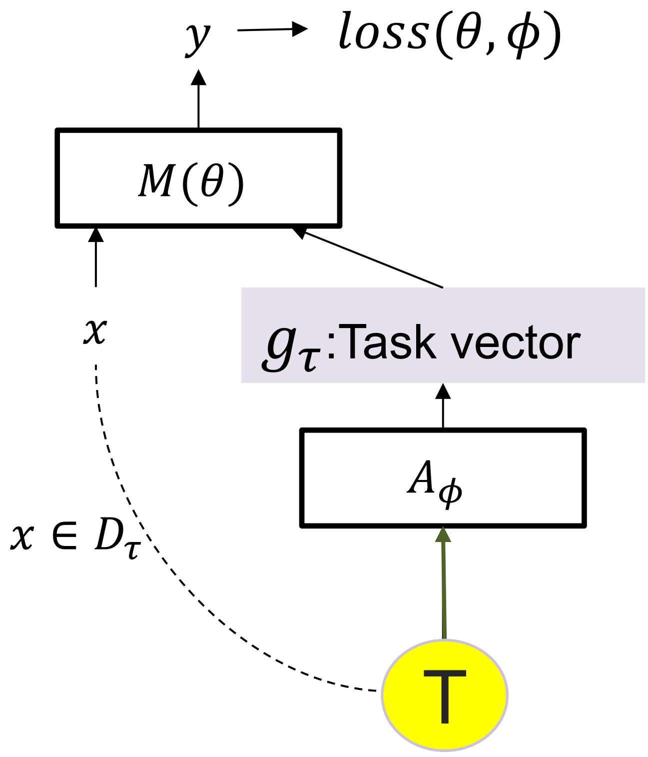


Mixture-of-Experts Meets Instruction Tuning: A Winning Combination for Large Language Models. ICLR 2024.

Shen, Hou, Zhou, Du, Longpre, Wei, Chung, Zoph, Fedus, Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, Denny Zhou

Adaptation with Task vectors

Another way to maintain task identities



Training:

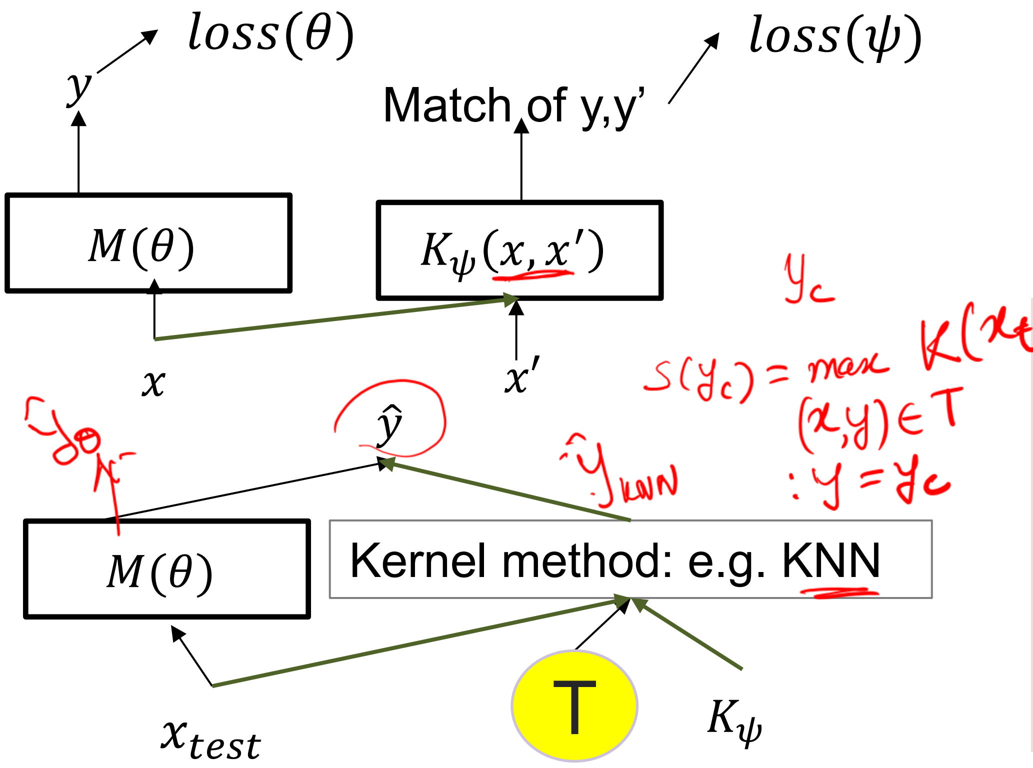
- Train $\mathbf{\Theta}, \phi$ jointly
- Requires task boundaries

Adaptation:

Use T to extract $g_t = A_\phi(T)$

$\hat{y} = M(x_{test}, g_t)$

Adaptation with Matching Networks



Training:

- Train Θ as usual
- Contrastive $K_\psi(x, x')$ so example pairs with same labels are close

Adaptation:

- Use kernel method to get predictions from T using $K_\psi(x, x')$, $x' \in T$.
- Blend with predictions from $M(x|\theta)$

- Labeled Memory Networks for Online Model Adaptation. In AAAI, 2018. Shiv Shankar and Sunita Sarawagi
- Speech-enriched Memory for Inference-time Adaptation of ASR Models to Word Dictionaries Mittal, Sarawagi, Jyothi. EMNLP 2023.
- Conditional Tree Matching for Inference-Time Adaptation of Tree Prediction Models. Varma, Awasthi, Sarawagi *ICML* 2023

An example matching-based adaptation for Text2SQL

Conditional Tree Matching for Inference-Time Adaptation of Tree Prediction Models. Varma, Awasthi, Sarawagi *ICML* 2023.

Text to SQL

Schema

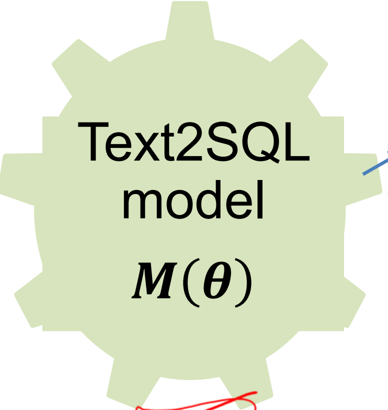
Department		
Id	Name	Description
14	Informatik	..
11	Software system	..
21	Engineer	…

Degree programs		
Id	Name	Dept id
1	PhD	14
2	Masters	21
3	Bachelors	21

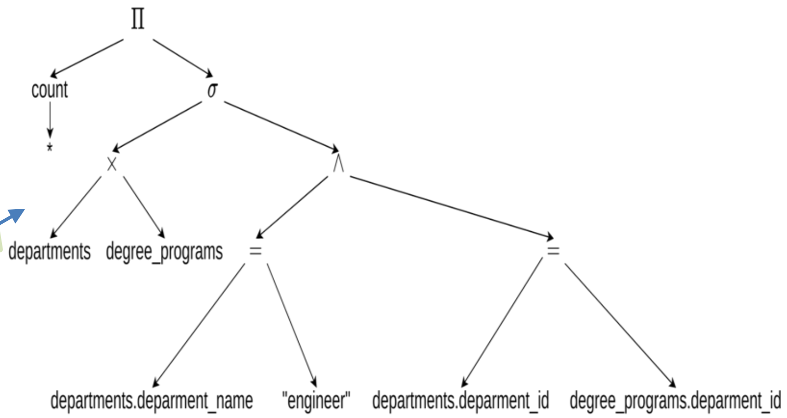
Question

How many degrees does the engineering department offer?

$\boldsymbol{x}$



Relational Algebra Tree representation of SQL



$\boldsymbol{y}$

Designing matching kernel: $K((x, y), (x', y'))$

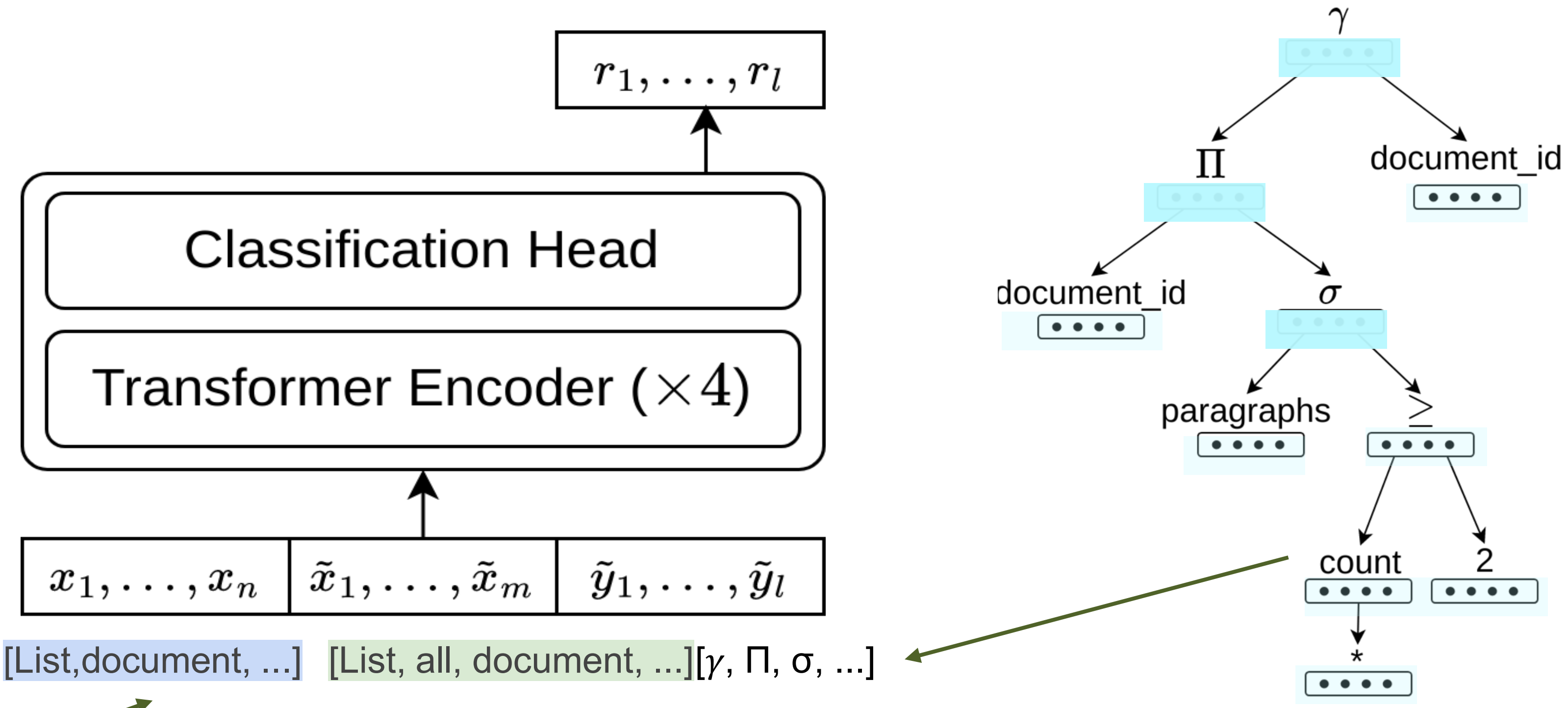
Two parts since y is structured.

- Node-level reasoning of the relevance of specific sub-trees of labelled examples to current question
  - Fine-grained sharing.
- Similarity based on alignment of respective trees → structurally informed

Fine-grained relevance scoring

Relevance score each subtree

Wrong nodes marked irrelevant



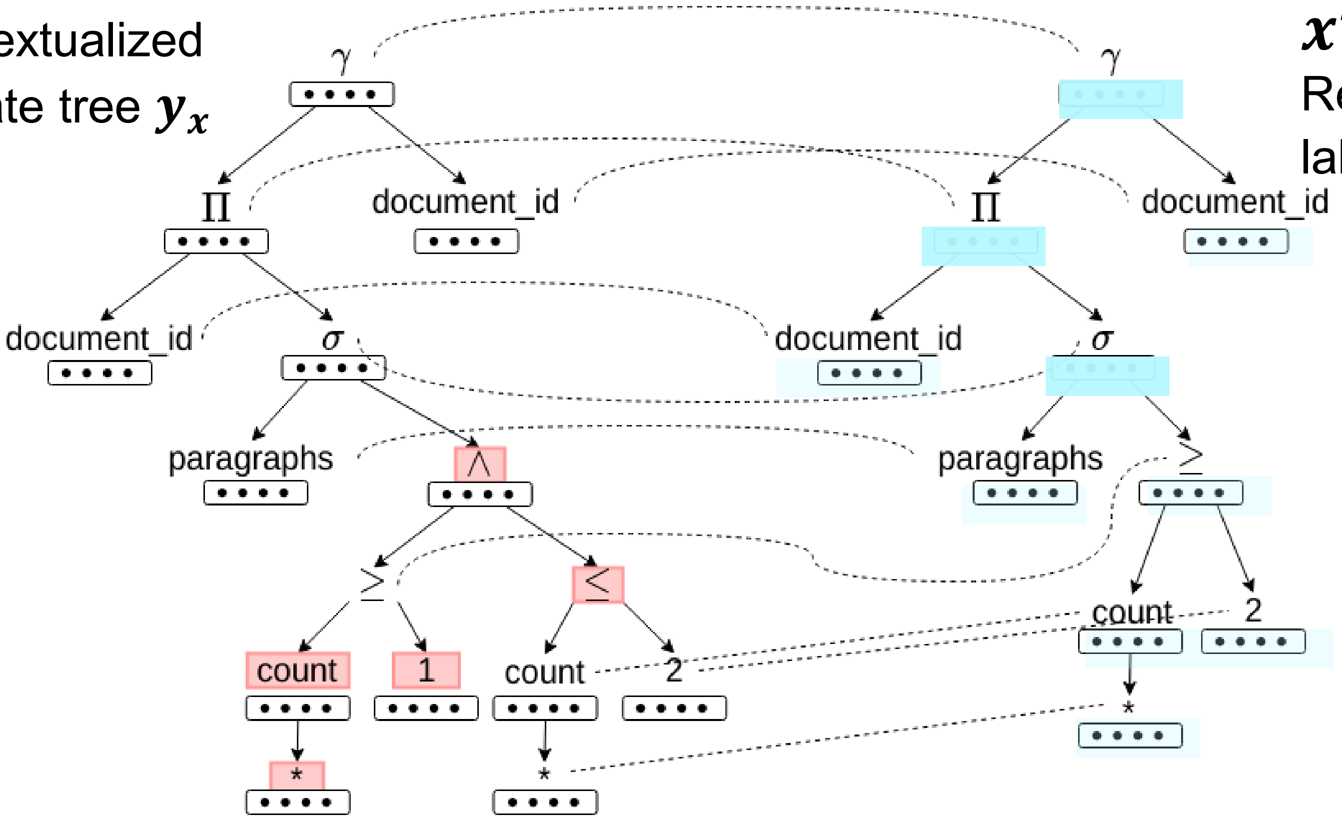
User Question $\boldsymbol{x}$

Labeled Question $\boldsymbol{x}'$, Tree $\boldsymbol{y}'$ from T

Alignment based relevance-weighted $K((x, y), (x', y'))$

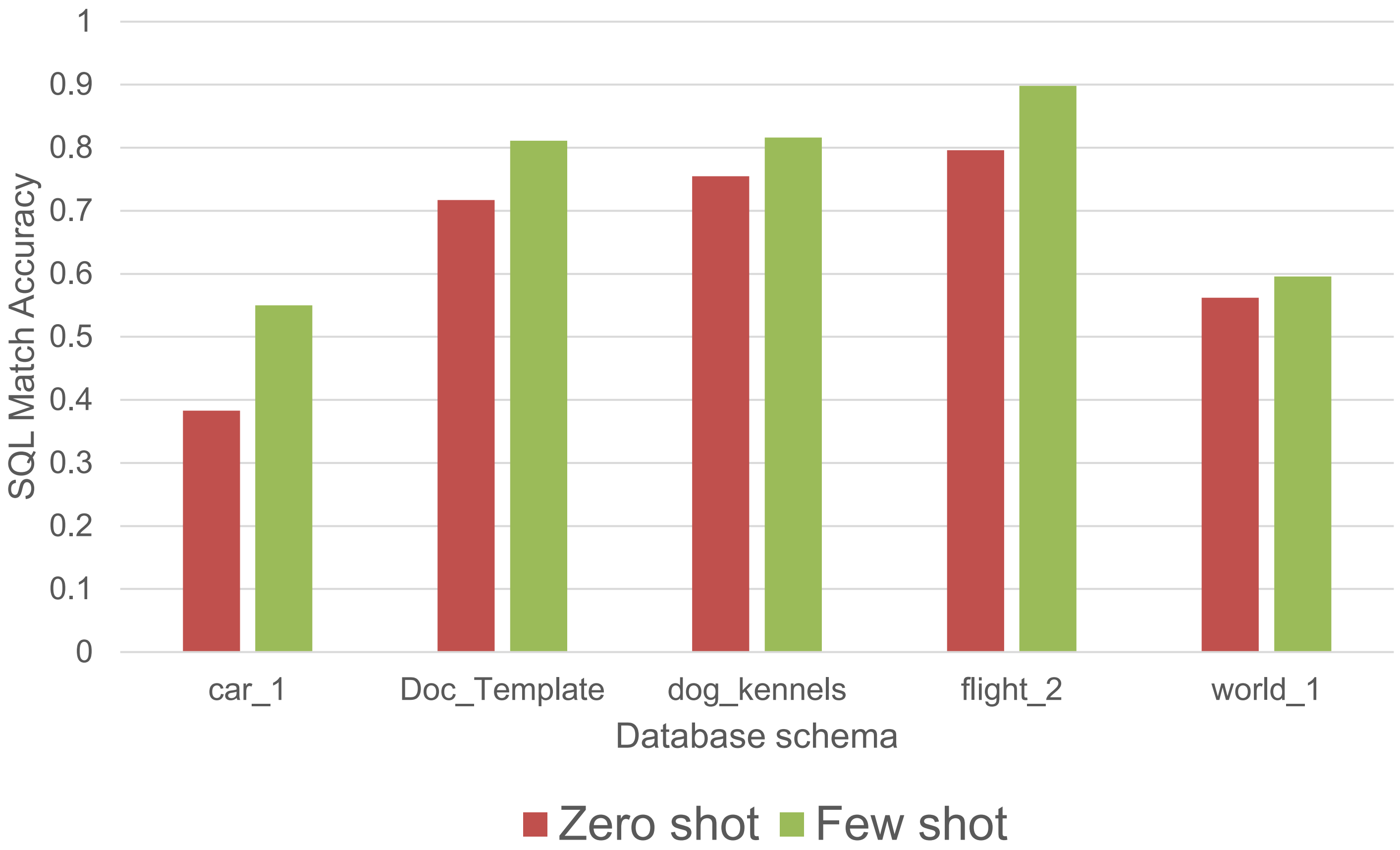
$\boldsymbol{x}$ Contextualized candidate tree $\boldsymbol{y}_{\boldsymbol{x}}$

$\boldsymbol{x}'$ Contextualized Relevance weighted labeled tree $\boldsymbol{y}'_{\boldsymbol{x}'}$

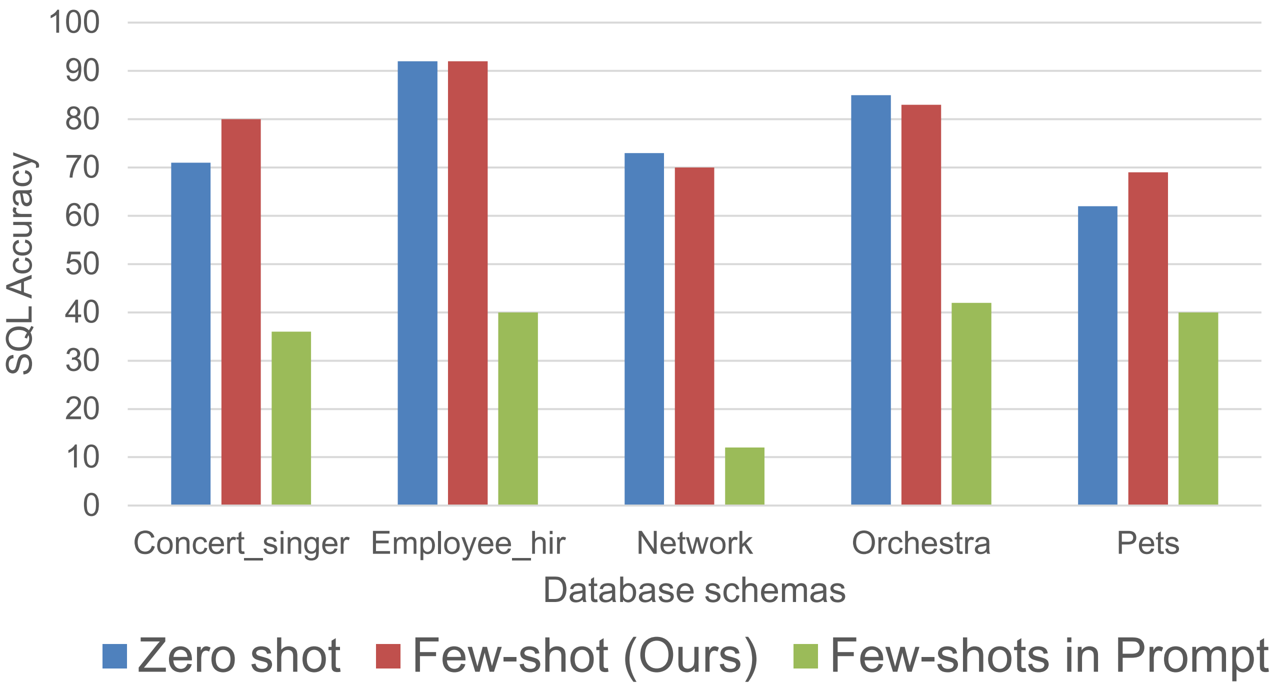


Conditional Tree Matching for Inference-Time Adaptation of Tree Prediction Models. Harshit Varma, Abhijeet Awasthi and Sunita Sarawagi In *ICML* 2023.

Accuracy after adaptation with few examples



Greater robustness to irrelevant examples



Reasoning about fine-grained relevance of few-shots is important for robustness

(small T)	Fine-tuning	MoE	Task vectors	Matching
Accuracy	✓✓	✓✓✓	✓	✓
Robustness	✓	✓✓	✓✓✓	✓✓
Efficiency				
Adapt	✓	✓	Online ✓✓✓	Online ✓✓✓
Test	✓✓✓	✓✓✓	✓✓✓	✓
Others		✓ Difficult to train		

Model adaptation methods

- Fine-tuning and its variants

- Mixture of experts and task-vectors

- Matching / memory augmented methods.

Each category explicitly planned for adaptation either when training $\mathrm{M}(\boldsymbol{\Theta})$ or adapting to T

Emergent Model Adaptation in LLMs

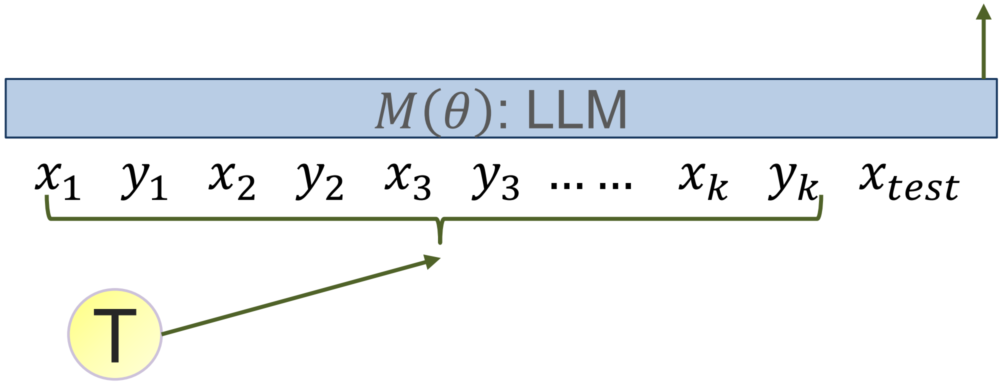
- Solution to model adaptation just emerges in the form of in-context learning!

- Models trained with next-token prediction loss on huge natural document collections can be adapted just with few labelled examples in-context..

Language models are few-shot learners. Brown et al NeurIPS 2020

Adaptation via In-Context Learning in LLMs

Predicted label $\hat{y} \sim M_T(x_{test}|\theta)$



Training: Default

Adaptation: Just a forward pass.

Language models are few-shot learners. Brown et al NeurIPS 2020

Examples

42

LLM: $M(\theta)$

2 ! 3 = 6, 3 ! 3 = 9, 5 ! 4 = 20, 6 ! 7 =

63

LLM: $M(\theta)$

2 ! 3 = 6, 3 ! 3 = 9, 5 ! 4 = 20, 9 ! 7 =

Why does ICL work?

- Hypothesis 1: Transformers implement gradient descent algorithm over IC examples

- Hypothesis 2: IC examples recognize tasks in pre-training e.g. via task vectors

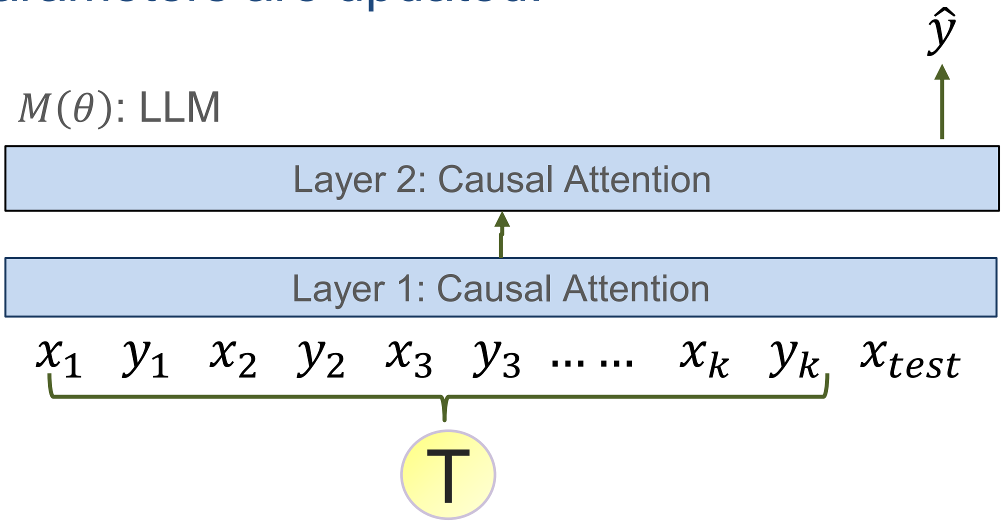
- Hypothesis 3: Self-attention implements matching-based adaptation via induction heads

We observe a parallel between these hypothesis and age-old methods of model adaptation.

Hypothesis 1: Transformers implement gradient descent over label loss during ICL

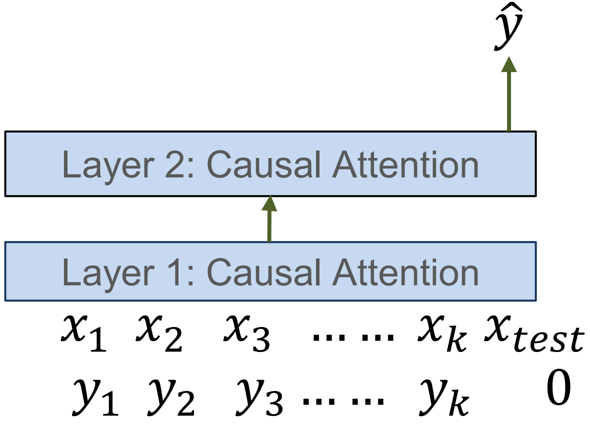
But how can GD be implemented in one forward pass over T?

- No obvious loss on IC examples, leave alone gradients on loss?
- No θ parameters are updated!



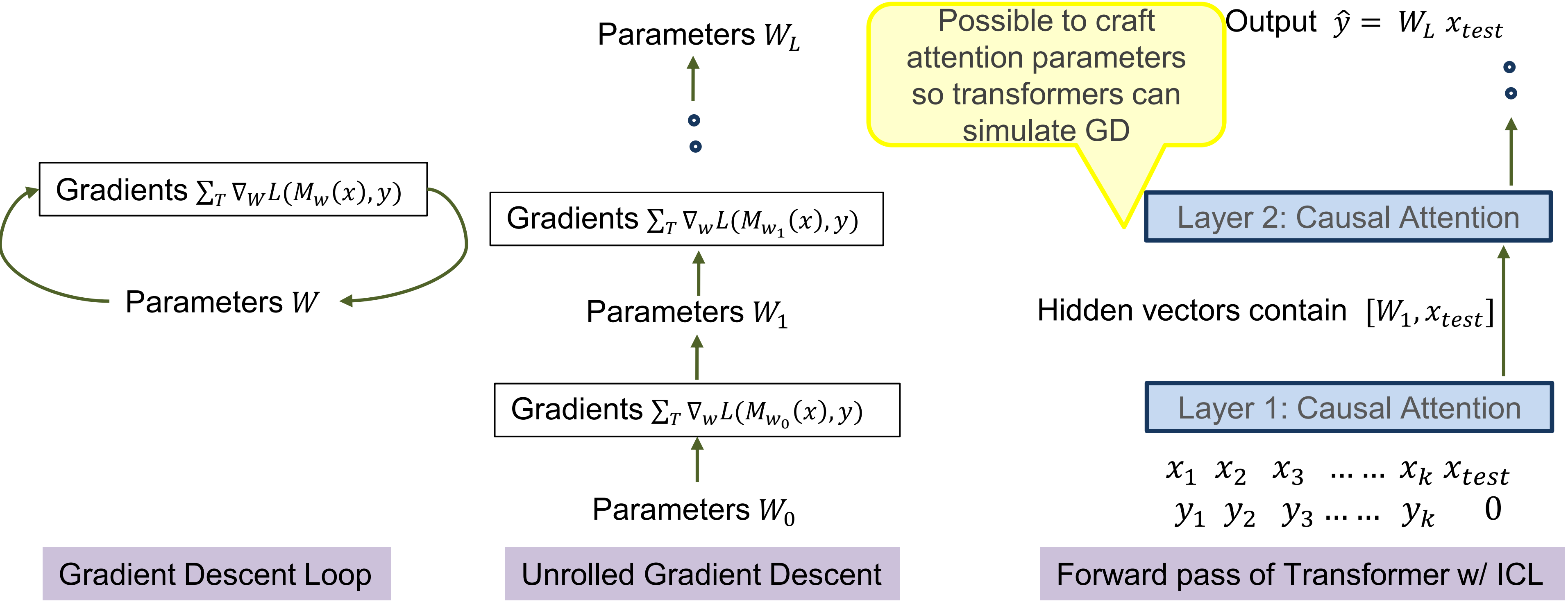
Assume

- Input stacked as [xi,yi]
- Task is linear regression
  - x_i is a real vector
  - $y_i = w_\tau x_i$
- Loss is square loss
  - $\min_{w} \sum_{(x_i,y_i)\in T} (y_i - w x_i)$^2
- Transformer attention is linear: no softmax.



What can transformers learn in-context? a case study of simple function classes. NeurIPS 2022. Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant.

Each layer implements a gradient step during forward pass of ICL



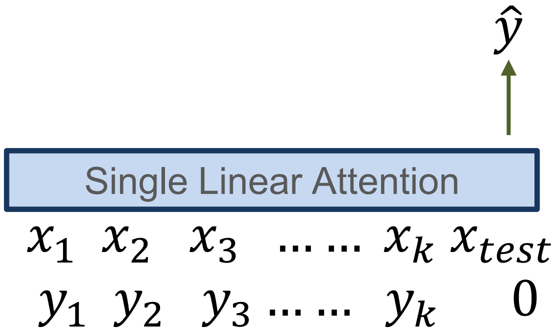
- Transformers learn in-context by gradient descent. Johannes Von Oswald et al. ICML 2023
- What learning algorithm is in-context learning? investigations with linear models. Ekin Akyurek et al. ICLR 2023

But can we learn these parameters that implement gradient descent using standard next-token loss?

Proved for special cases

Special case I: One-layer linear Transformer

- Optima of training loss leads to transformer parameters that implement single pre-conditioned gradient descent step.

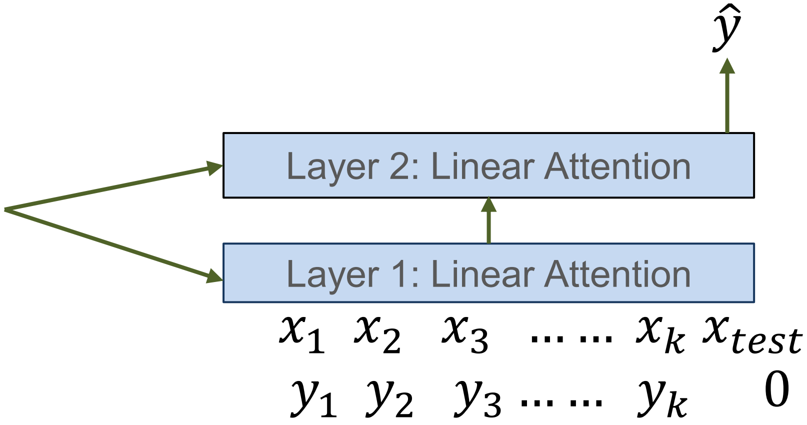


- Trained transformers learn linear models in-context. JMLR 2024. Ruiqi Zhang, Spencer Frei, Peter L. Bartlett
- Transformers learn to implement preconditioned gradient descent for in-context learning. NeurIPS 2023. Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, Suvrit Sra

Proved for special cases

Special case II: Linear Multi-layer looped Transformer

Same parameters θ in the two layers of the transformer



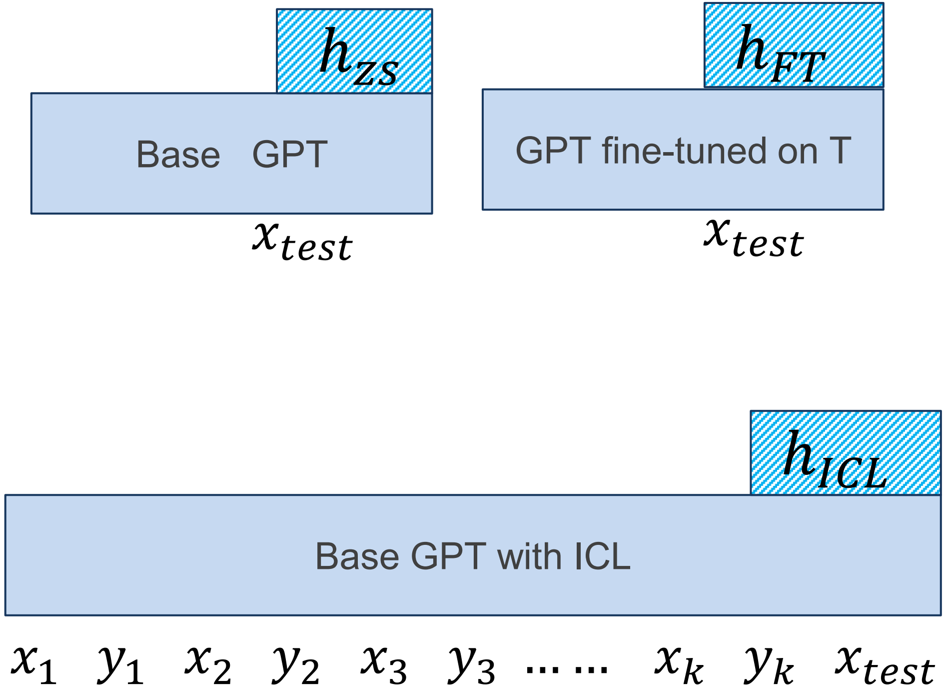
But these results are for numerical regression tasks, and that too on special transformers trained for this task. Do not generalize to text and NLP

Can looped transformers learn to implement multi-step gradient descent for in-context learning? ICML 2024.
Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar

What about text and pre-trained LLMs?

Dai et al (2023) support GD hypothesis for ICL for text and pre-trained LLMs

$$\boxed{h_{FT} - h_{zs} \sim h_{ICL} - h_{zs}}$$



Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. ACL 2023. Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei.

ICL performs Gradient descent hypothesis is refuted..

Even an untrained transformer has high similarity between corresponding vectors while showing no capability for ICL

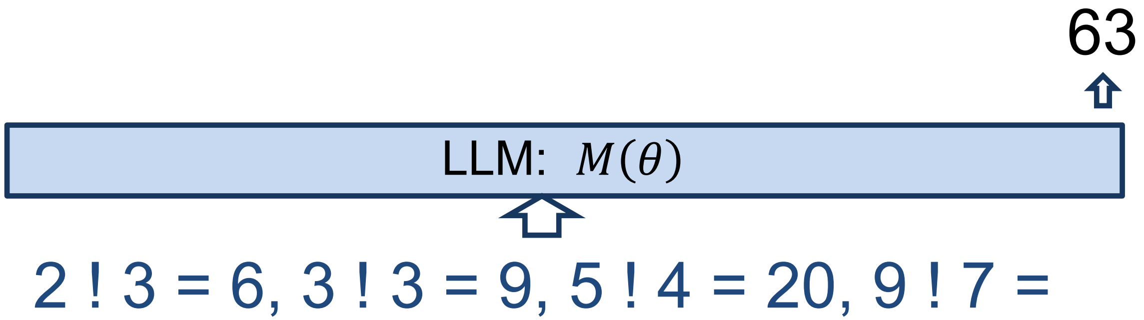
In-context learning and gradient descent revisited. NAACL 2024. G Deutch, N Magar, T Natan, G Dar.

Others: Li et al ACL 2024, Shen et al ICML 2024

Hypothesis 2:

ICL identifies task from the pre-training set

ICL does task selection

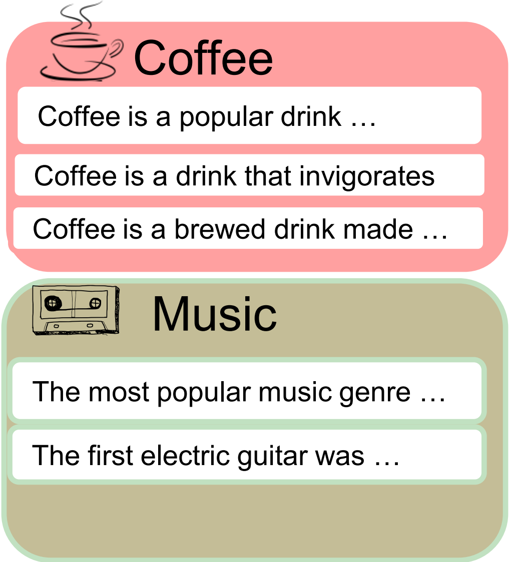


LLMs provide the correct answer by recognizing this as a multiplication problem. They do not really learn to do multiplication from these examples.

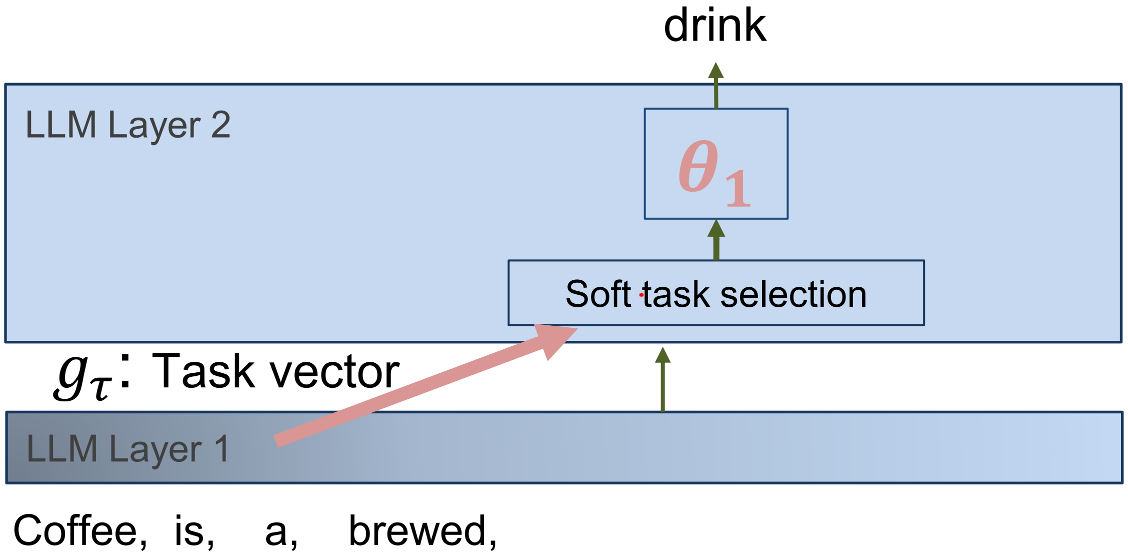
How does task selection capability emerge in LLMs after pre-training with next token loss?

ICL does Task Selection

Pretraining documents: mix of related topics



Prefix of sentence recognizes topic as latent vectors g_τ → soft select task → recall words from related documents seen earlier.



- An explanation of in-context learning as implicit Bayesian inference. ICLR 2022. Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma
- The learnability of in-context learning. NeurIPS 2023. Noam Wies, Yoav Levine, and Amnon Shashua

ICL goes beyond task selection

Large LLMs do learn novel input-output mappings

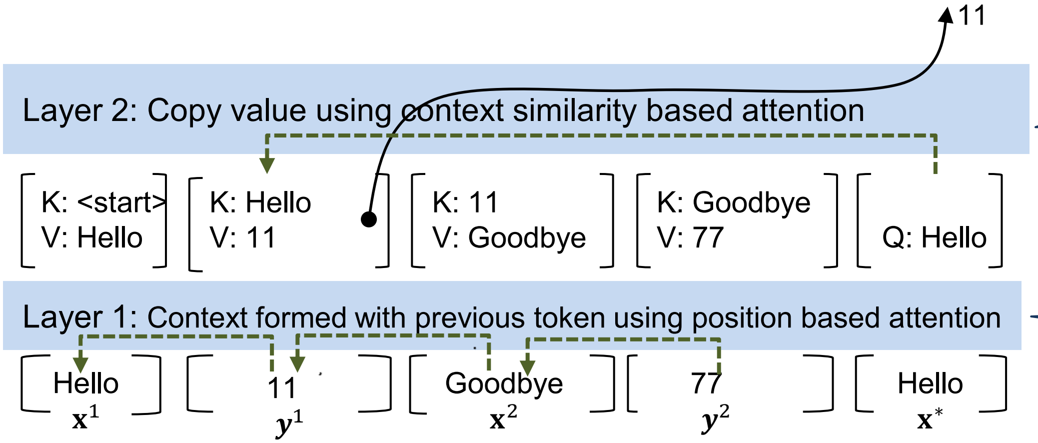
- Replacing positive/negative with foo/bar or flipping labels works for sentiment classification
- Unseen key-value associative mappings are learned
- Totally synthetic formal Markovian languages are learned

- What do language models learn in context? the structured task hypothesis. ACL 2024. Jiaoda Li, Yifan Hou, Mrinmaya Sachan, and Ryan Cotterell.
- Why larger language models do in-context learning differently? ICLR 2024. Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang

Hypothesis 3: Self-attention implements matching-based adaptation

ICL= Matching with Induction circuits

Induction circuits created over two layers



Layer 2: Matching with

- Key = copied previous token,
- Value = current token,
- Query = current token

Layer 1's attention copies previous token as key.

In-context learning and induction heads. Olsson et al 2022

The mechanistic basis of data dependence and abrupt learning in an in-context classification task. ICLR 2024. Gautam Reddy

But why do transformer parameters orient to create in-context induction heads during next-token loss pre-training?

Pre-training data contains repeated phrases

Corpus contains co-occurring phrases in similar templates.

care. We will ***estimate*** infant and neonatal mortality rates again in ***2015*** to see if this trend continues and, if so, to assess how it can ***be*** reversed. Infant mortality in ***2013*** was

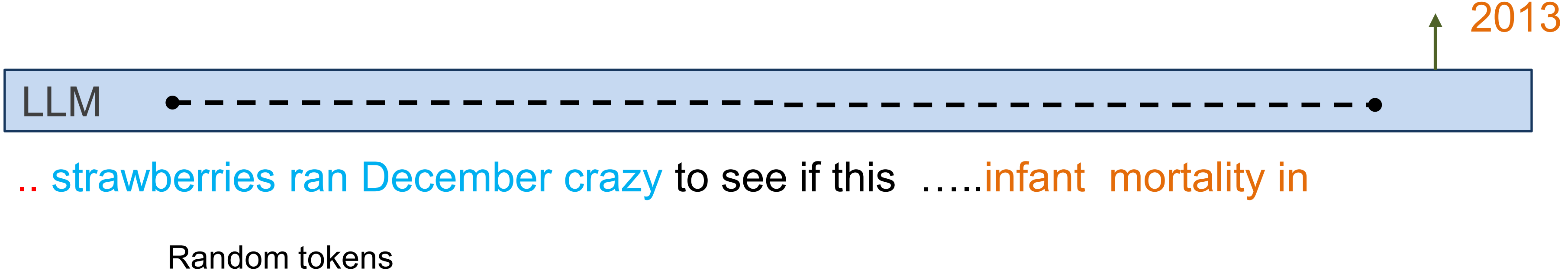
.. neonatal mortality rates again in 2015 to see if this …..infant mortality in

Parallel structures in pre-training data yield in-context learning. Chen et al. In ACL, 2024. Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, He He

Pre-training data contains repeated phrases

Destroy repetitions from context during training causes ICL to drop by 50% as against 2% random drops

care. We will ***estimate*** infant and neonatal ~~mortality rates again in ***2015***~~ to see if this trend continues and, if so, to assess how it can ***be*** reversed. Infant mortality in ***2013*** was



Parallel structures in pre-training data yield in-context learning. Chen et al. In ACL, 2024. Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, He He