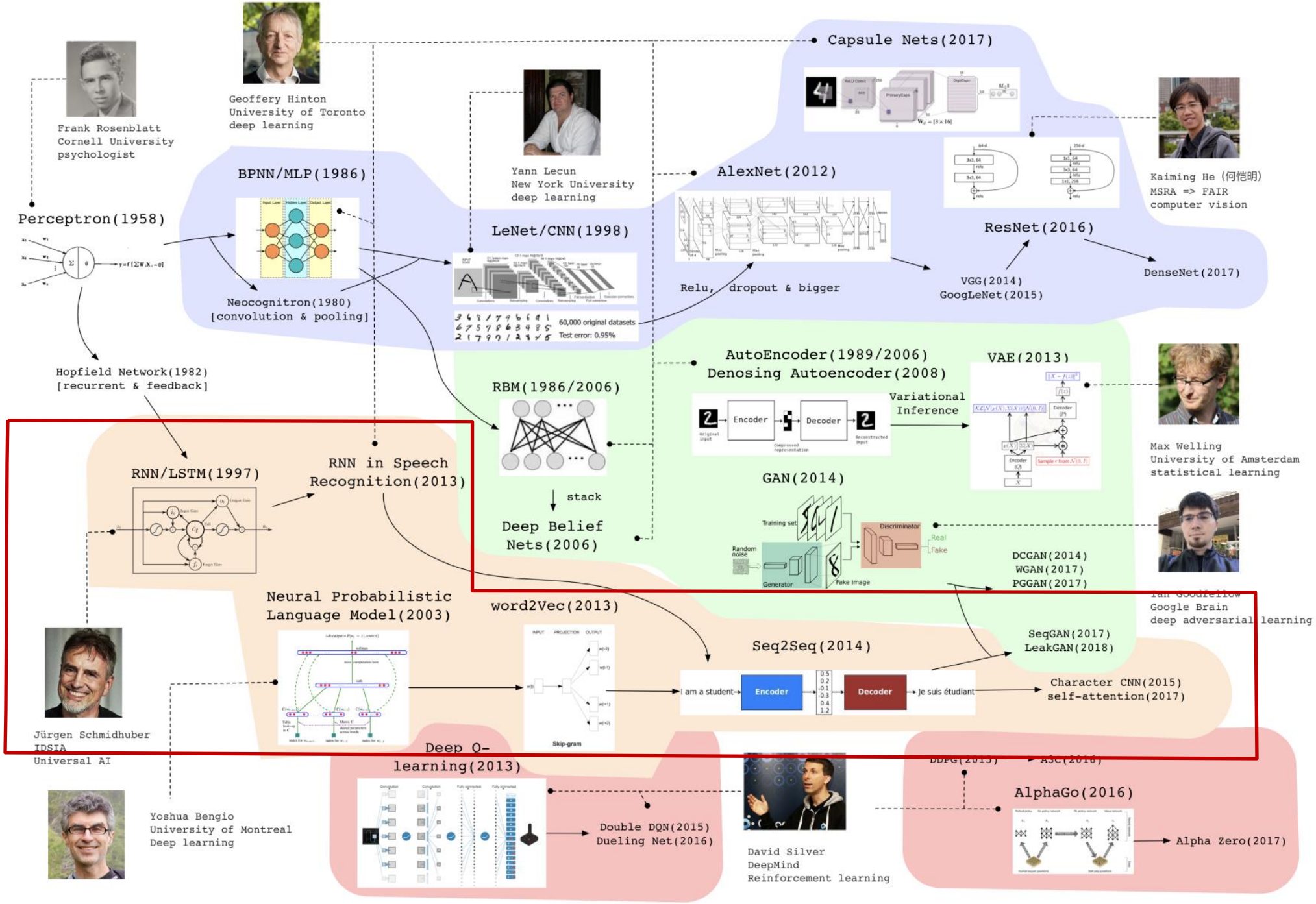
In-Context Learning & Emergent Abilities

Jie Tang
Department of Computer Science & Technology
Tsinghua University

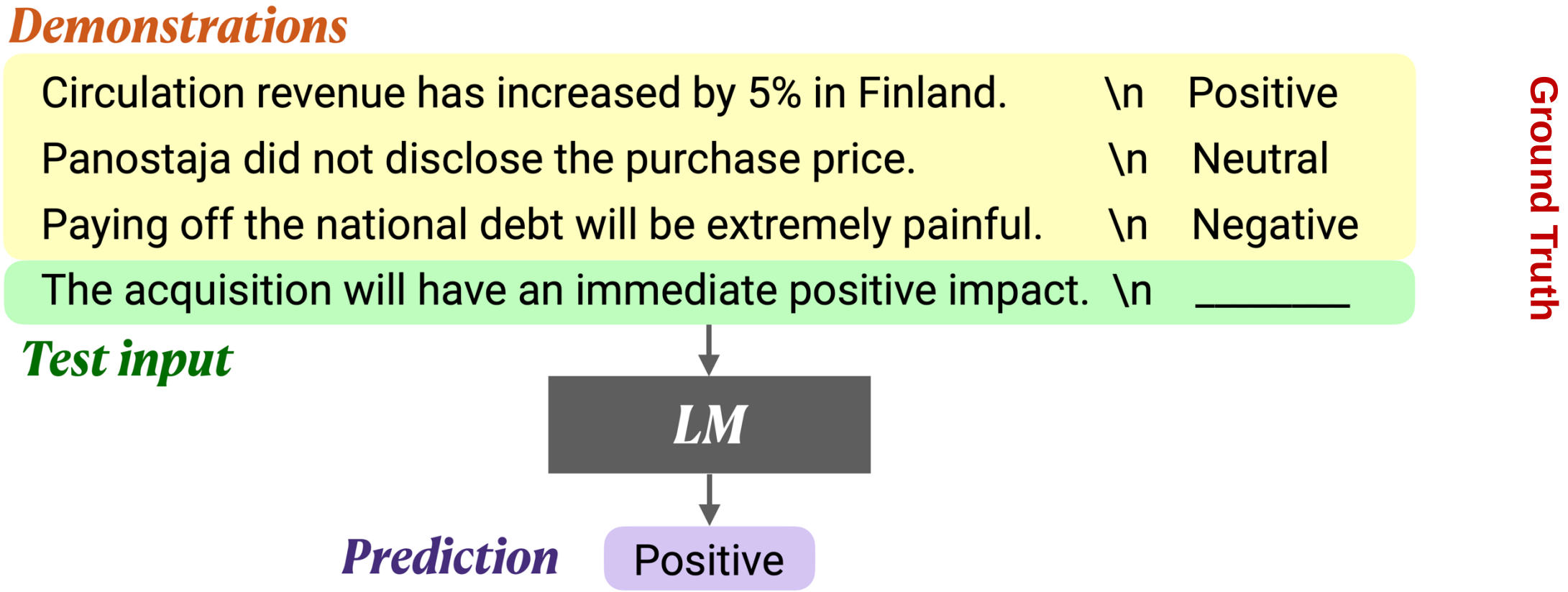

Capsule Nets(2017)
Geoffery Hinton
University of Toronto
deep learning
Frank Rosenblatt
Cornell University
psychologist
Yann Lecun
New York University
deep learning
BPNN/MLP(1986)
AlexNet(2012)
Kaiming He (何恺明)
MSRA => FAIR
computer vision
Perceptron(1958)
LeNet/CNN(1998)
ResNet(2016)
Neocognitron(1980)
[convolution & pooling]
60,000 original datasets
Test error: 0.95%
Relu, dropout & bigger
VGG(2014)
GoogLeNet(2015)
DenseNet(2017)
Hopfield Network(1982)
[recurrent & feedback]
AutoEncoder(1989/2006)
Denosing Autoencoder(2008)
VAE(2013)
Encoder
Decoder
Variational
Inference
Max Welling
University of Amsterdam
statistical learning
RBM(1986/2006)
stack
Deep Belief
Nets(2006)
GAN(2014)
Training set
Random
noise
Generator
Discriminator
Real
Fake
Fake image
DCGAN(2014)
WGAN(2017)
PGGAN(2017)
Google Brain
deep adversarial learning
RNN/LSTM(1997)
RNN in Speech
Recognition(2013)
Neural Probabilistic
Language Model(2003)
word2Vec(2013)
Skip-gram
Seq2Seq(2014)
I am a student
Encoder
Decoder
Je suis étudiant
SeqGAN(2017)
LeakGAN(2018)
Character CNN(2015)
self-attention(2017)
Jürgen Schmidhuber
IDSIA
Universal AI
Yoshua Bengio
University of Montreal
Deep learning
Deep Q-
learning(2013)
Double DQN(2015)
Dueling Net(2016)
David Silver
DeepMind
Reinforcement learning
AlphaGo(2016)
Alpha Zero(2017)

Summary

	Long-Term Dependency	Parallelization	Computation	Flexibility for Length	Gradient Flow	Complexity
RNN	Poor, due to recurrent structure	Limited, due to sequential processing	Slow, especially for long sequences	Flexible for input, fixed for output	Problems with vanishing/exploding gradients	Simpler, fewer hyperparameters
LSTM	Better, still rooms to improve	Limited, due to sequential processing	Slow, especially for long sequences	Flexible for input, fixed for output	Better gradient flow with less vanishing	Complex, many hyperparameters
CNN	Not inherent	Highly parallelizable	Fast due to parallelizable operations	Less flexible, typically requires fixed-size inputs	Stable gradient flow	Complex, many hyperparameters
Bahdanau Attention	Good with attention mechanism	Limited, due to sequential processing	Slow, especially for long sequences	Flexible for both input and output	Better gradient flow with less vanishing	Simpler, fewer hyperparameters
Transformer	**Good with attention mechanism**	**Highly parallelizable**	**Fast due to parallelizable operations**	**Flexible for both input and output**	**Stable gradient flow**	**Simpler, fewer hyperparameters**

In-context Learning

- In-context Learning: *A is X; B is Y;* Please answer: *C is __ ?*



In-context Learning

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

Positive

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

Finance

In-context Learning

- Pretraining & Finetuning are costly when adapting to a specific task
  - And impossible to cover all usecases.
- In-context learning enables LM to ***learn during inference***.

Context-induced Agent actions

Question	Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?
Thought 1	The question simplifies to "The Simpsons" character Milhouse is named after who. I only need to search Milhouse and find who it is named after.
Action 1	Search[Milhouse]
Observation 1	Milhouse Mussolini Van Houten is a recurring character in the Fox animated television series The Simpsons voiced by Pamela Hayden and created by Matt Groening.
Thought 2	The paragraph does not tell who Milhouse is named after, maybe I can look up "named after".
Action 2	Lookup[named after]
Observation 2	(Result 1 / 1) Milhouse was named after U.S. president Richard Nixon, whose middle name was Milhous.
Thought 3	Milhouse was named after U.S. president Richard Nixon, so the answer is Richard Nixon.
Action 3	Finish[Richard Nixon]

LLM Learning vs. Traditional ML

- 针对特定任务，利用标注好的数据 $\{(x_i,y_i)\}_{i=1...N}$，学习映射函数$f(\cdot)$

$$f(x) \rightarrow y$$

Recall traditional ML…

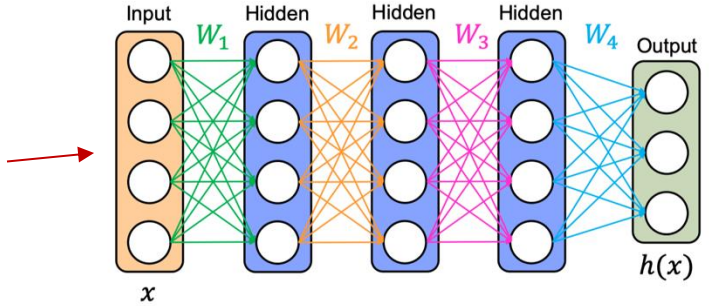
- 基于海量无标注数据学习模型model(·)

$$model(f(x) \rightarrow x)$$

- 应用于不同任务时给定例子$\{(a_i, b_i)\}_{i=1...k}$，自动迁移学习到的映射能力，进行预测

$$f(model(f(x) \rightarrow x), (a, b), a') \rightarrow b'$$

可以理解为在新样本(a,b)的带动下，部分网络被激活



当$k<<N$且上述针对类似任务预测精度相当的时候，我们说模型具有了“举一反三”的迁移能力

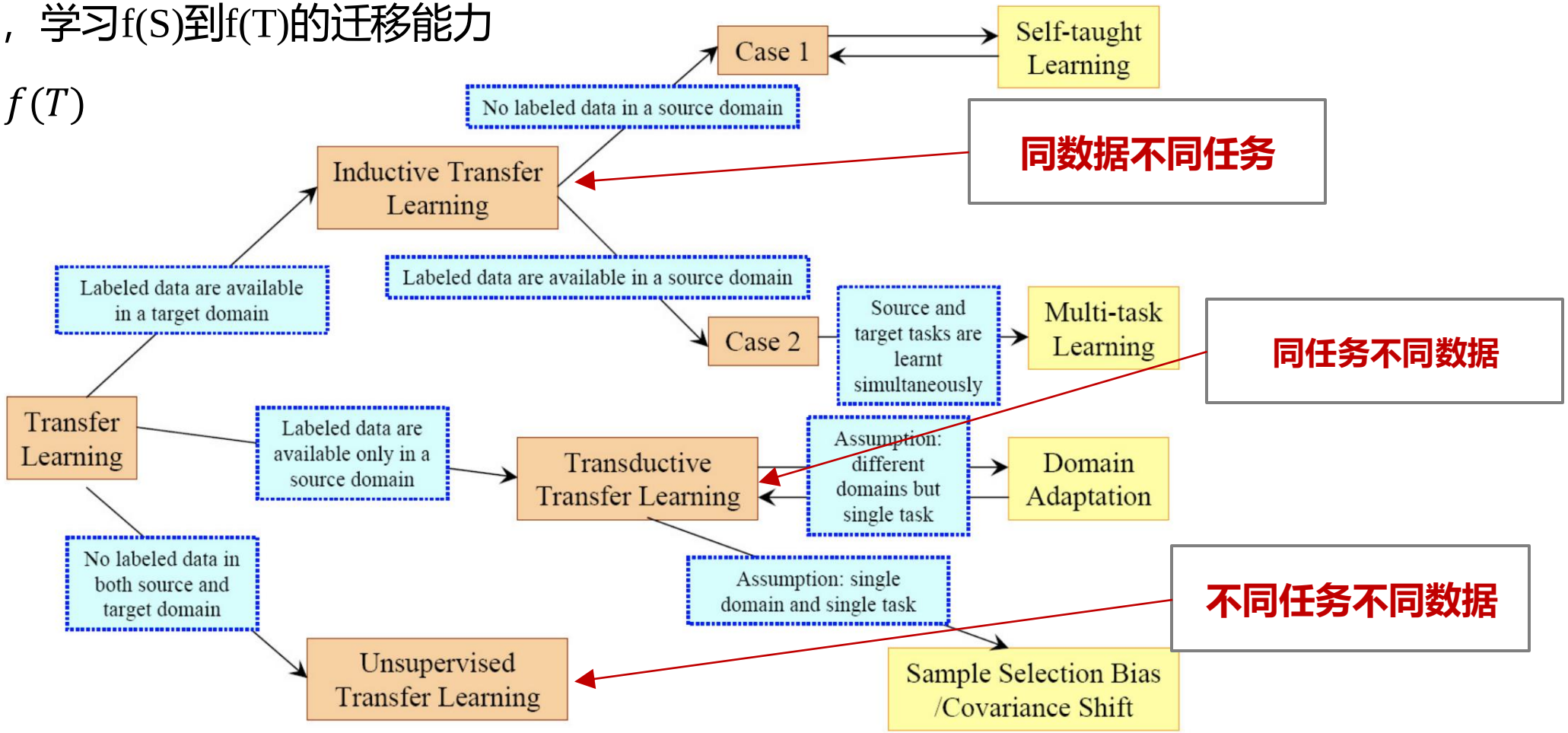
cost(f(下游任务)) x #下游任务 > cost($model(\cdot)$) + cost(迁移学习) x #下游任务

大模型的临界点

Transfer Learning

- 两个数据S、T，学习f(S)到f(T)的迁移能力

$$f(S) \rightarrow f(T)$$

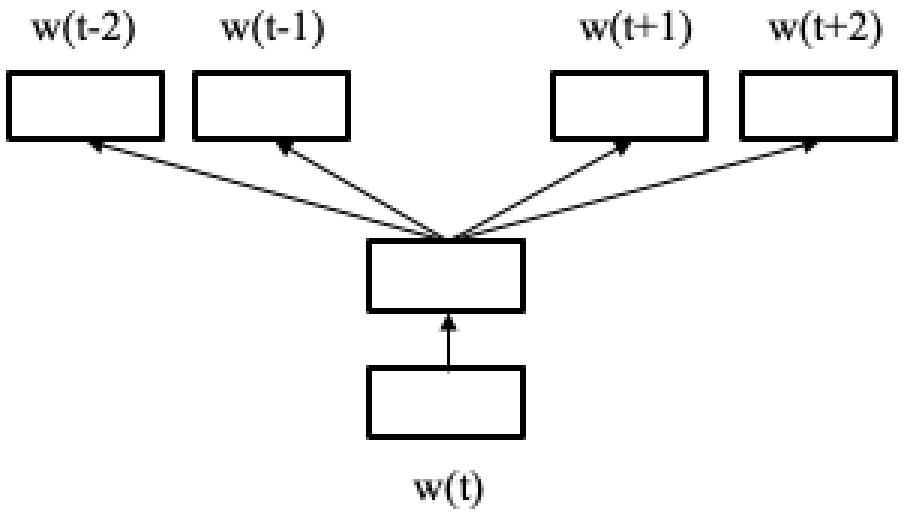


cost(f(S)) + cost(f(S)→f(T)) >> cost(f(S)) + cost(f(T))

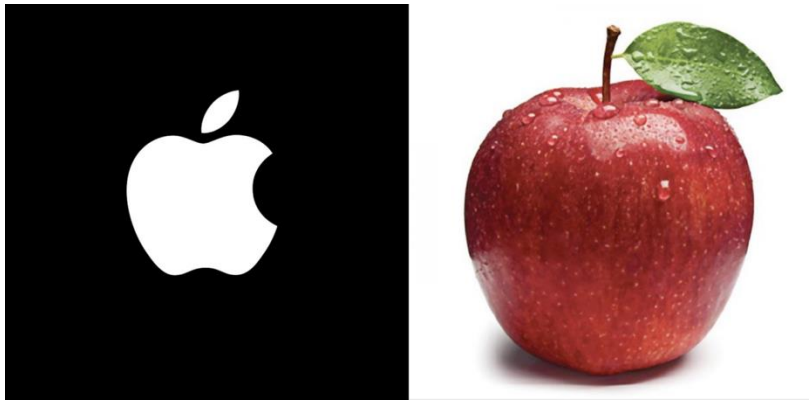
迁移学习的临界点

Representation Learning

- "You shall know a word by the company it keeps." —John Rupert Firth
- Learn static word representations



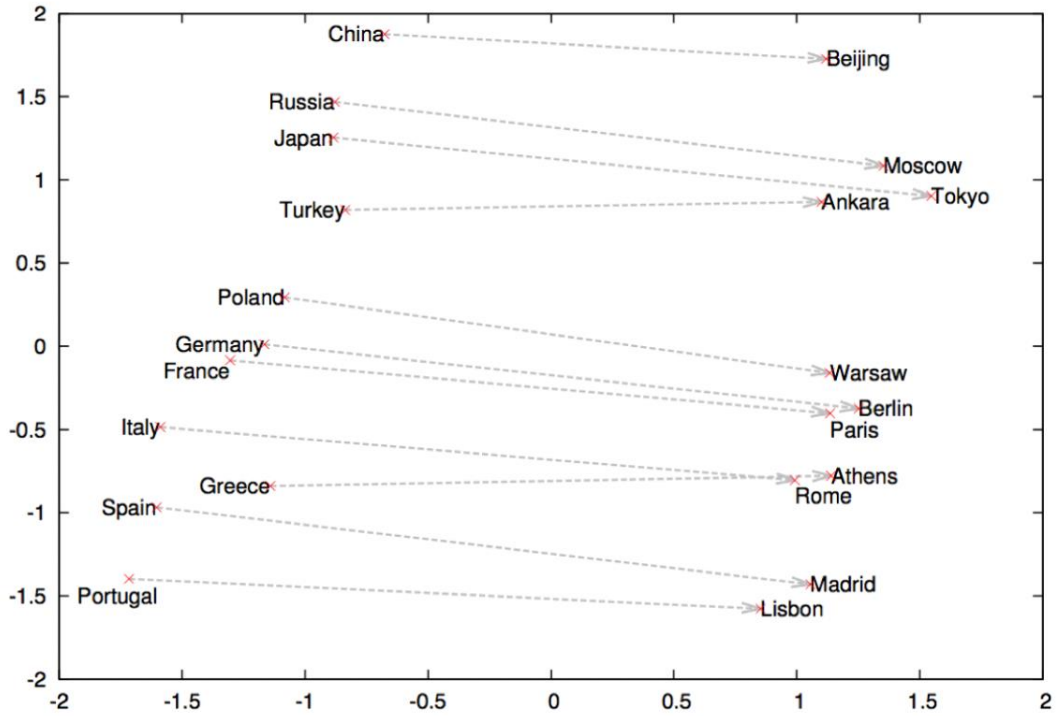
0.8	0.2	0.3	…	0.0	0.0



Apple vs. Apple?

Transfer based Representation Learning

- 基于单词级关系的迁移

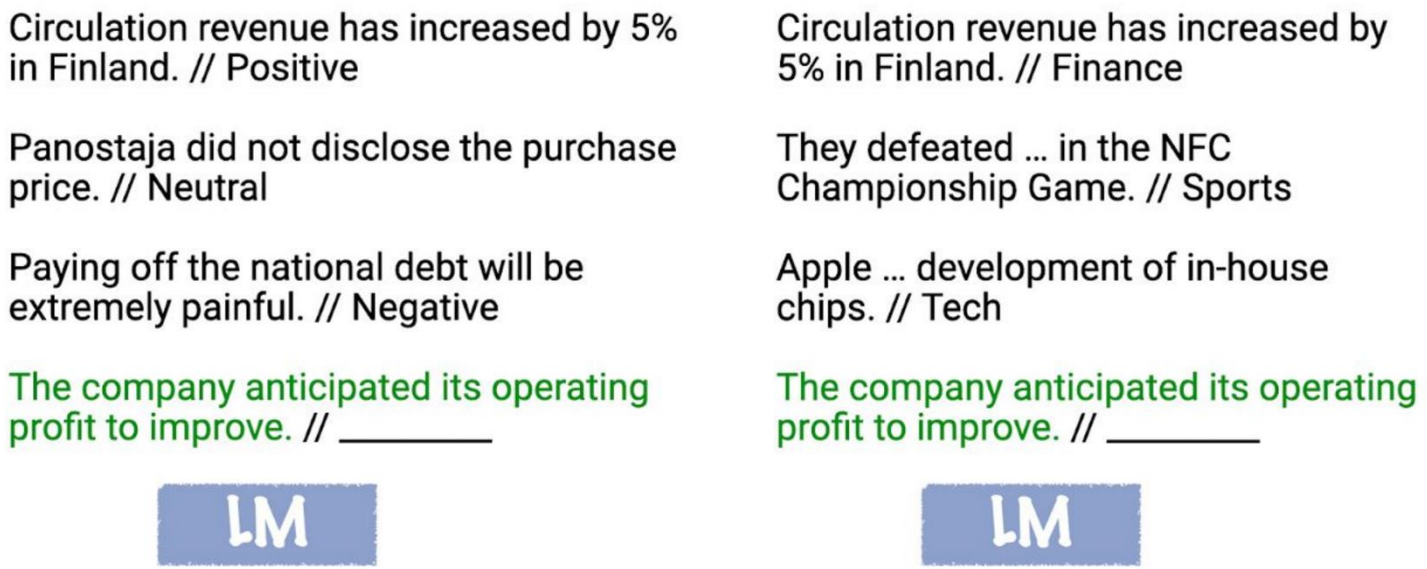


$$W(\text{"China"}) - W(\text{"Beijing"}) \approx W(\text{"Japan"}) - W(\text{"Tokyo"})$$

In-Context Learning

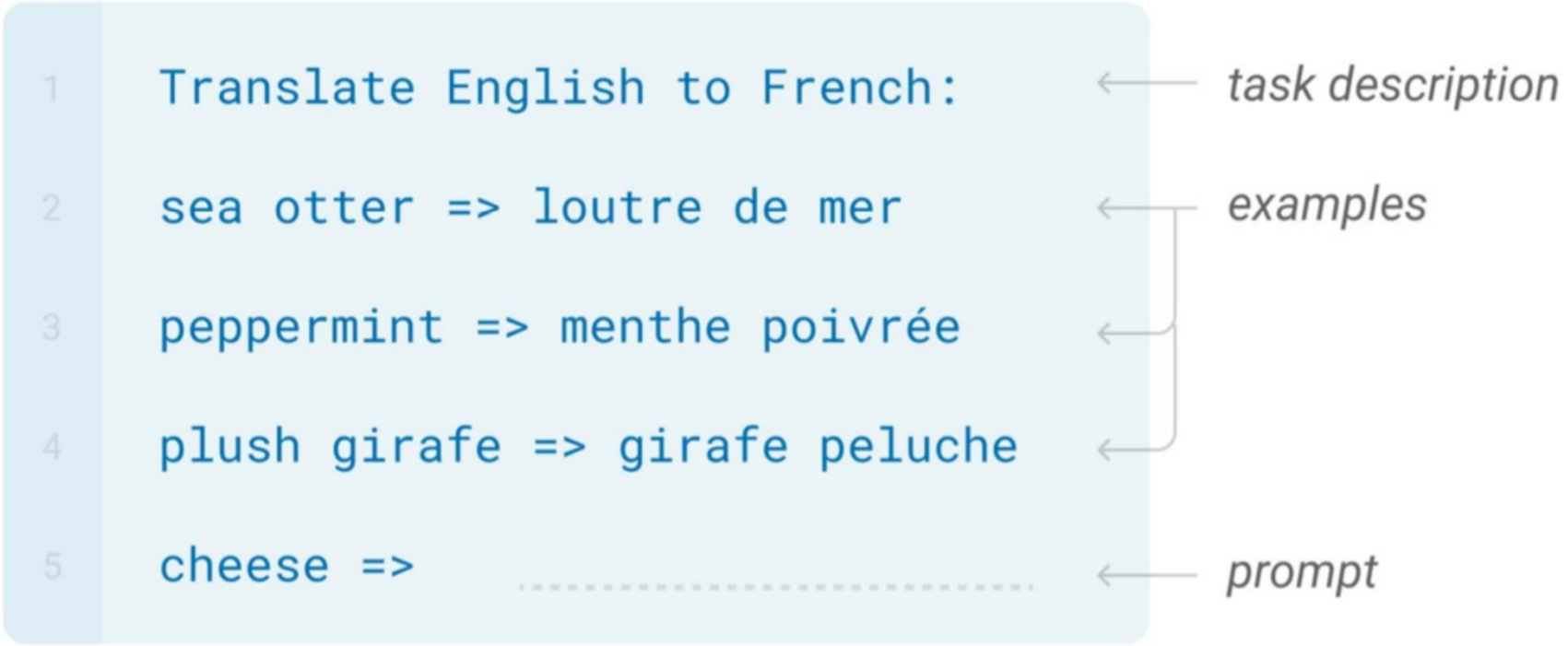
- Previous Deep Learning
  - Learning from supervised signals
  - Through parameter gradient descent

- In-Context Learning:
  - **Learn from Analogy**



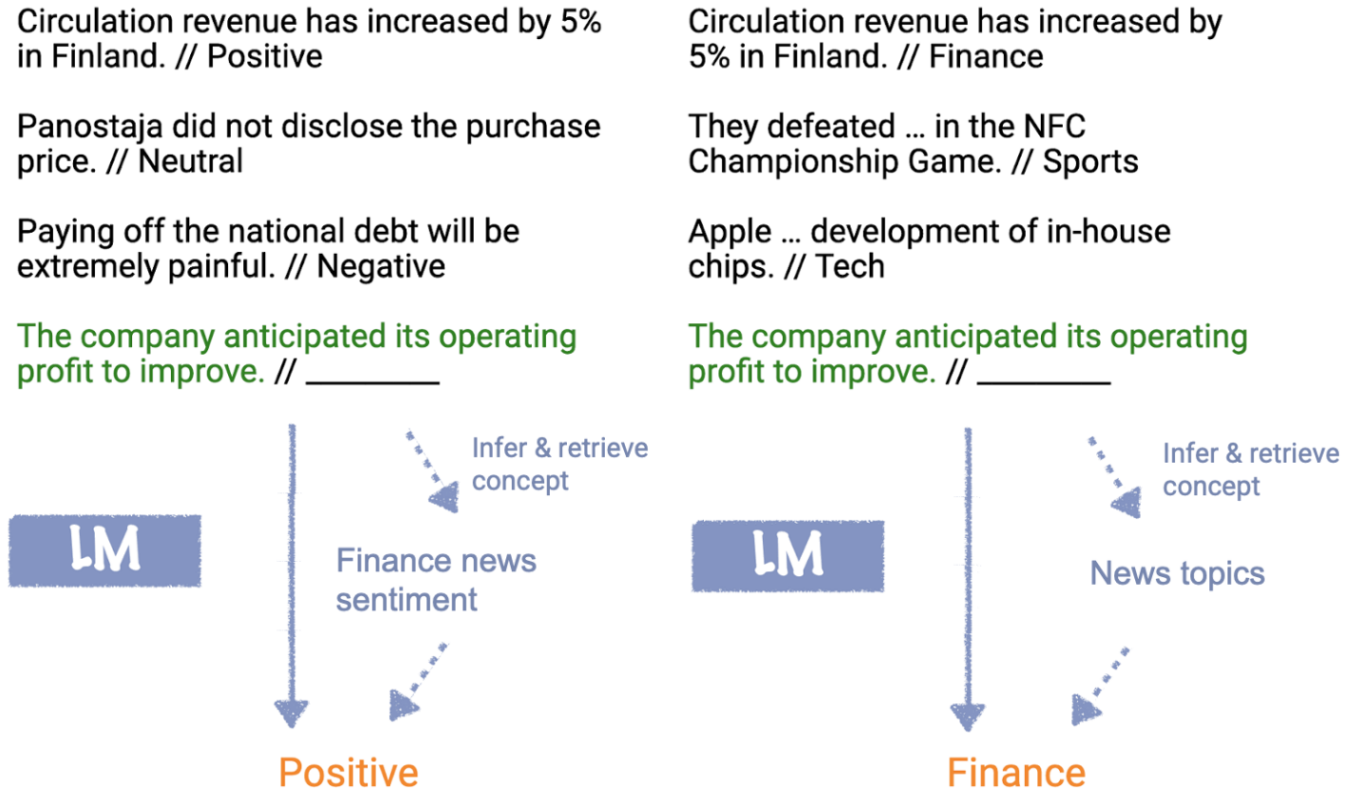
In-Context Learning

- Task description (Optional)
- Examples (Demonstrations)
- Prompt (Question / Problem)



Why In-Context Learning Works

- *A mysterious emergent behavior in large language models (Intuitively)*
- Language model uses the in-context learning prompt to “locate” a previously learned concept to do the in-context learning task



Context Construction

- How to select examples
  - Correctness of Examples
  - Varying the Number of Examples
  - Impact of input-output mapping
  - Bias in context examples
- How to design examples
  - Sensitive to the format of examples
  - Chain of thought

Capacity of LLMs

- In-context Learning: *A is X; B is Y;* Please answer: *C is __ ?*

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

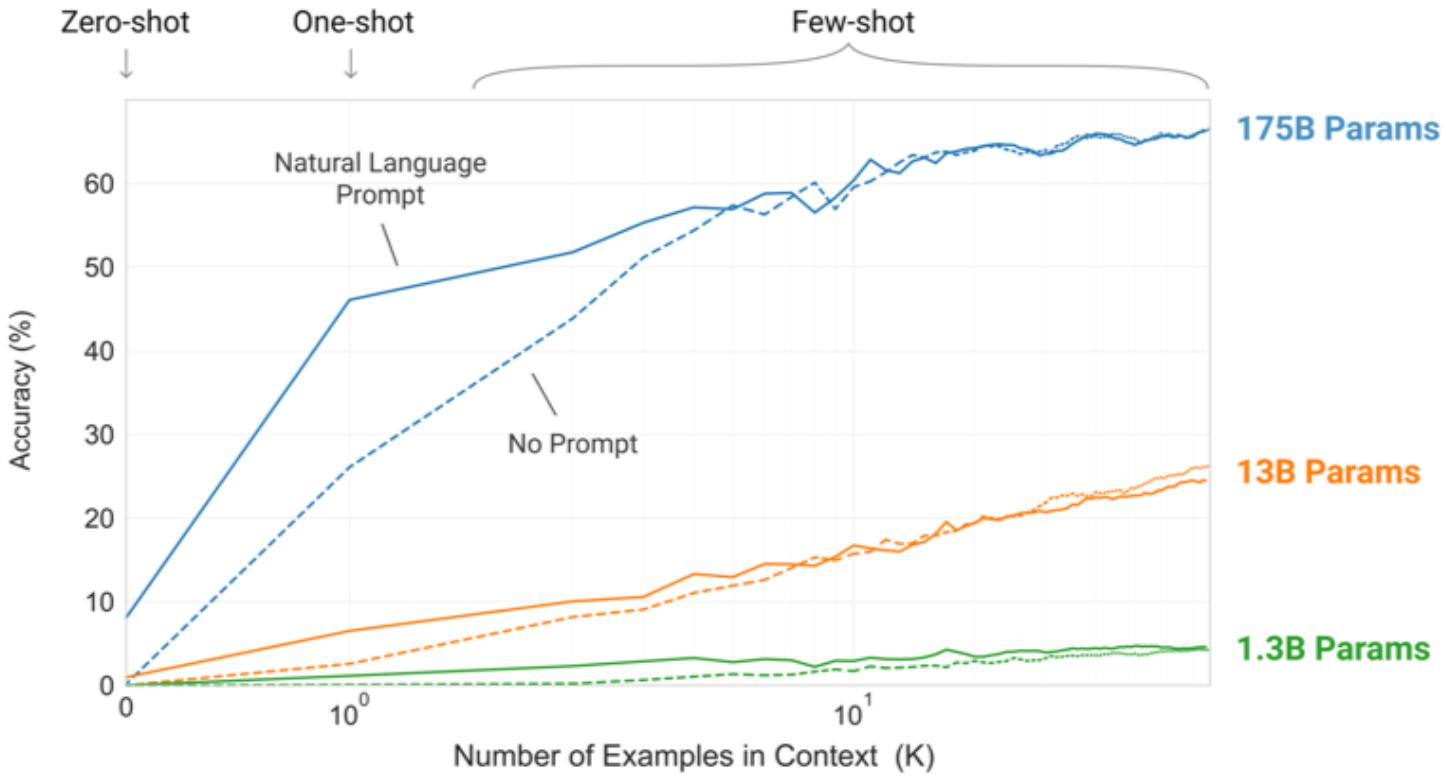
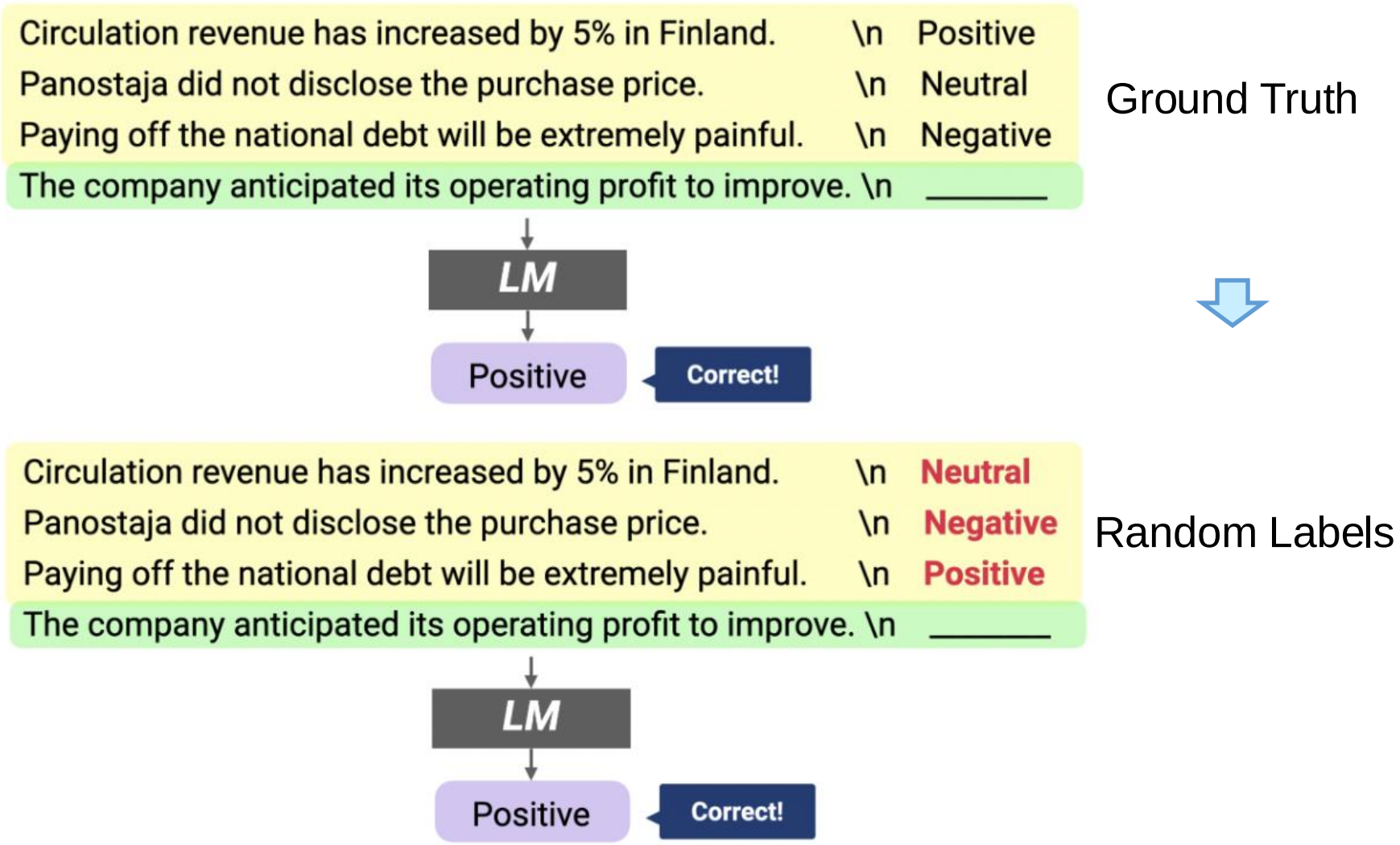


figure credit: GPT-3

Capacity of LLMs

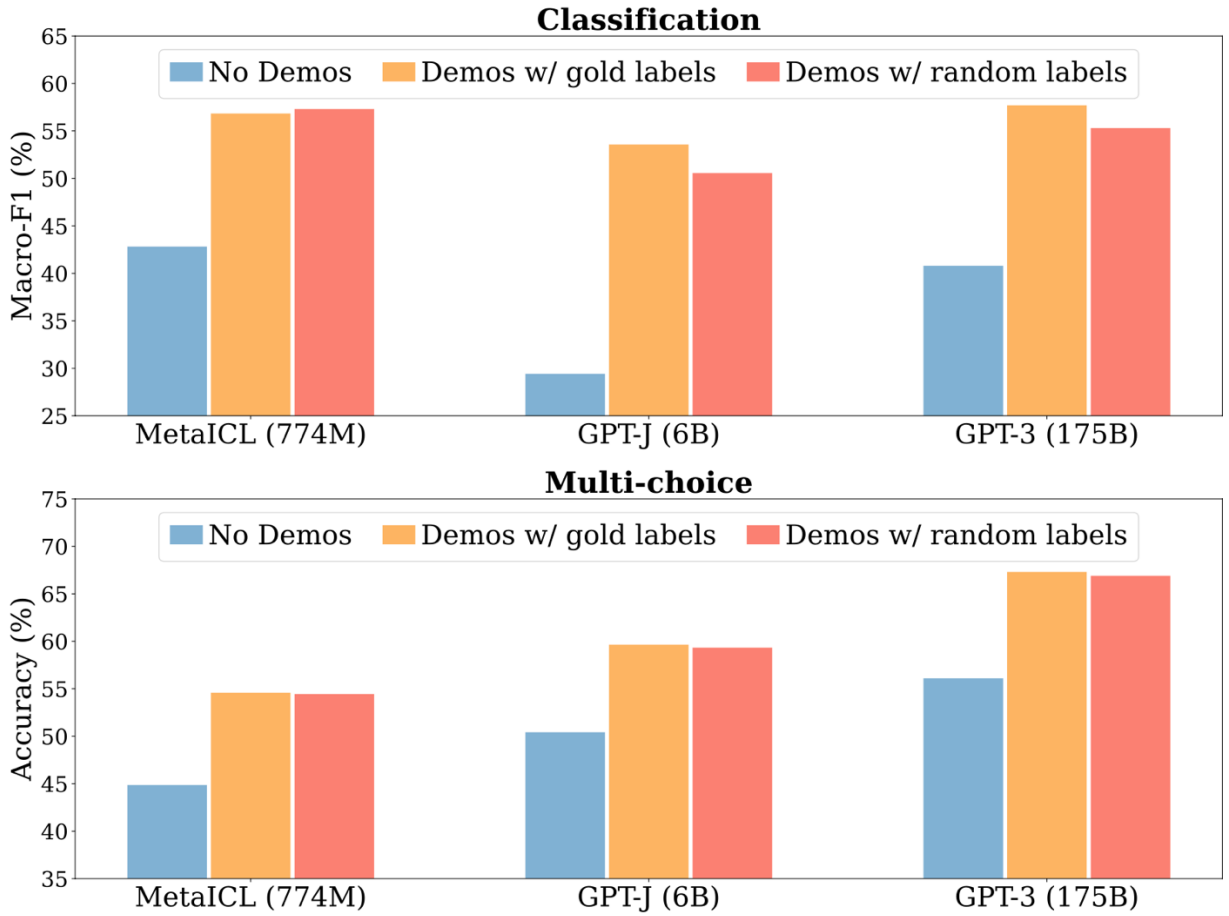
- In-context Learning: *A is X; B is Y;* Please answer: *C is __ ?*



UW, Meta, Allen4AI. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Prediction using three LMs with varying size.

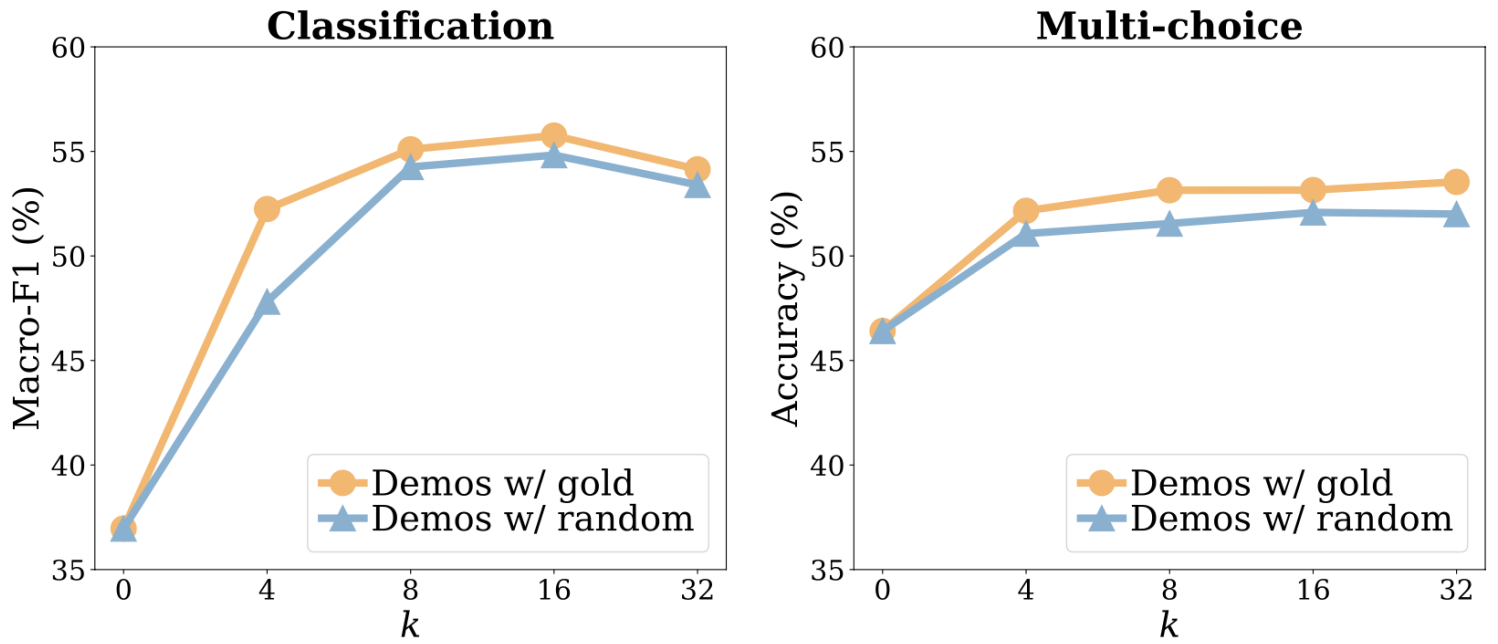
- In-context learning performance drops only marginally when labels in the demonstrations are replaced by random labels.



UW, Meta, Allen4AI. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Is the result consistent with varying k?

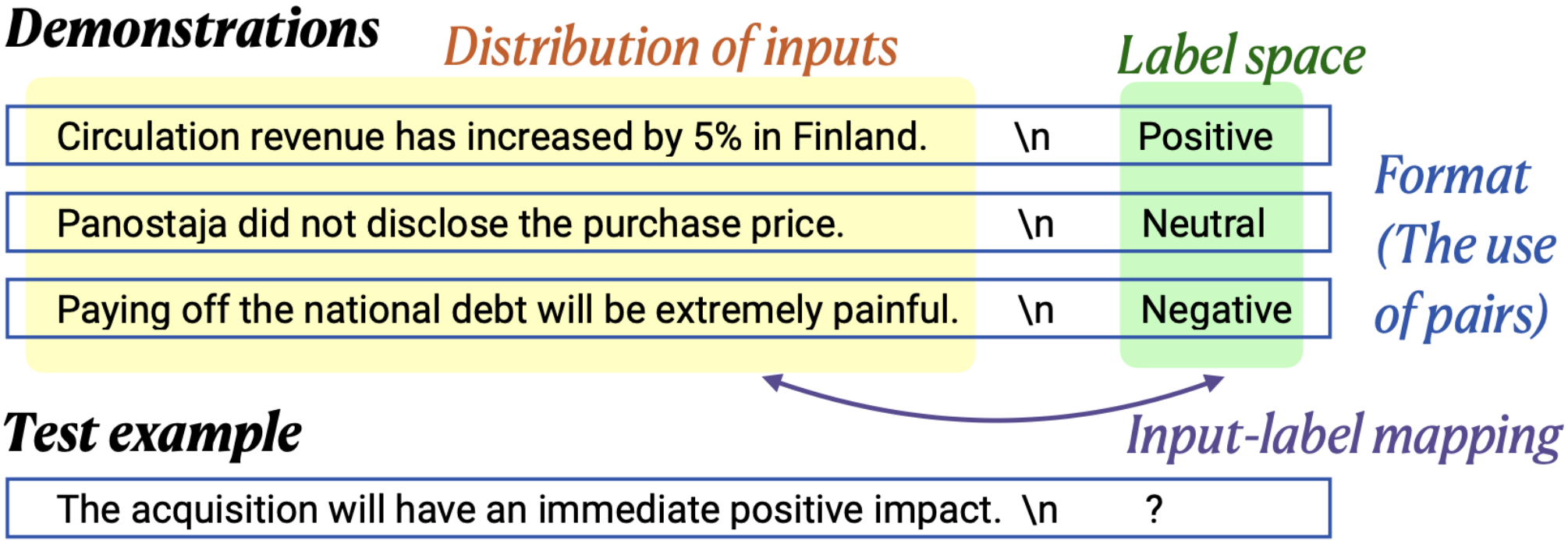
- First, using the demonstrations significantly outperforms the no demonstrations method even with small k (k = 4)
- ICL is beneficial mainly for supervising the label space, and other components are easier to recover from the data



UW, Meta, Allen4AI. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Signals for In-Context Learning

- The input-label mapping, the distribution of the input text, the label space, the format



The Input-Output Distribution

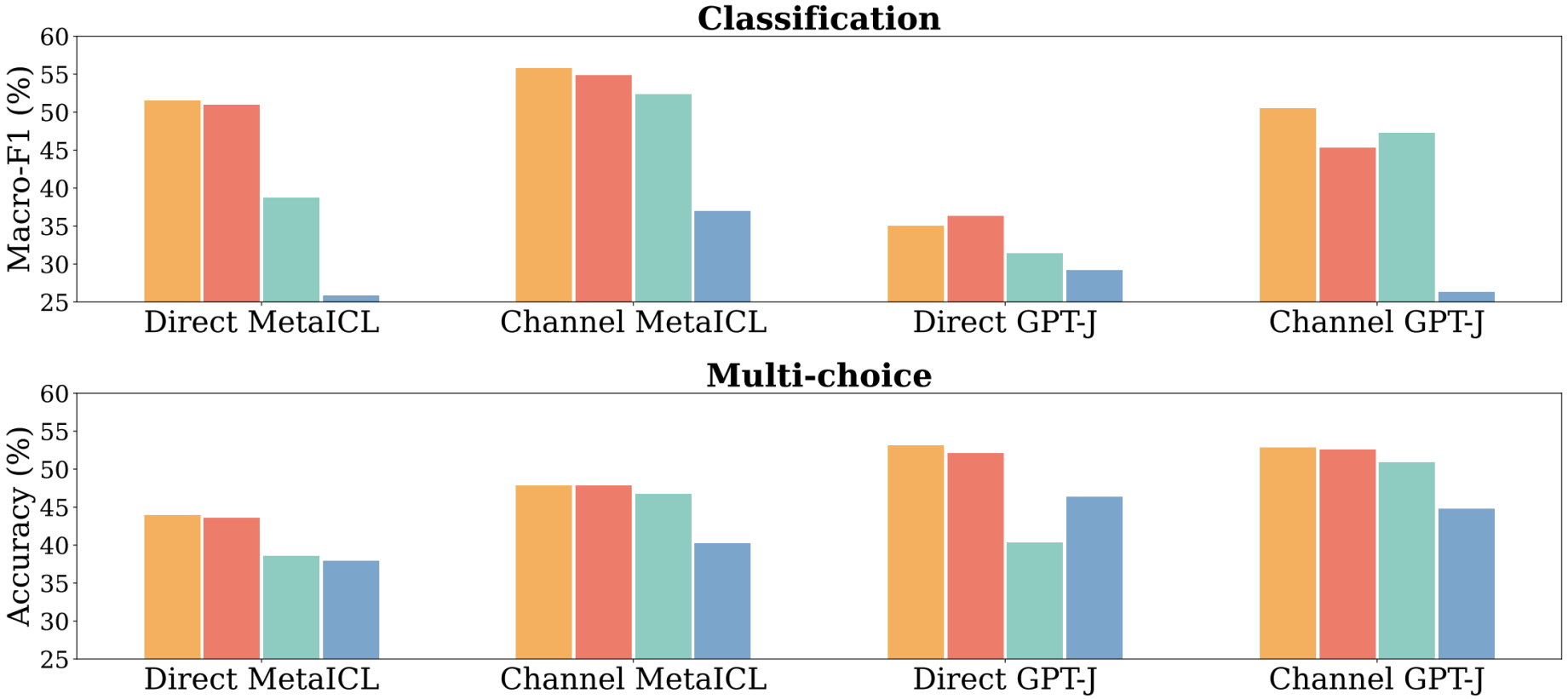
- Derivation of input examples
  - Sampled from downstream training data related to the *prompt*
  - Randomly sampled from an external corpus,

Circulation revenue has increased by 5% in Finland.	\n	Positive
Panostaja did not disclose the pruchase price.	\n	Neutral
Paying off the national debt will be extremely painful.	\n	Negative
The company anticipated its operating profit to improve.	\n	_____

Color-printed lithograph. Very good condition.	\n	Positive
Many accompanying marketing … meaning .	\n	Neutral
In case you are interested in learning more about ...	\n	Negative
The company anticipated its operating profit to improve.	\n	_____

The Input-Output Distribution

- In-distrubution examples vs. Out-of-distribution examples

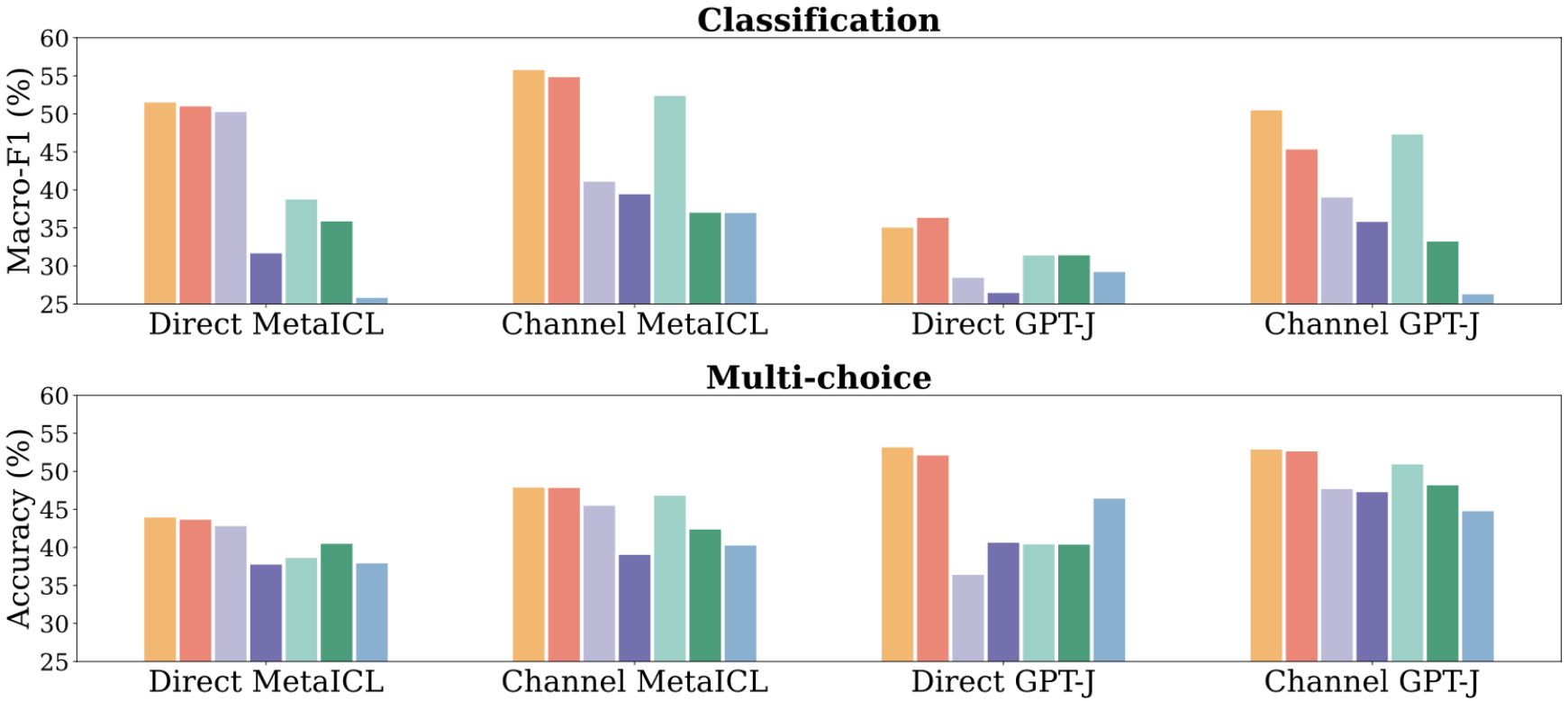


	F	*L*	*I*	*M*
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
Random English words	✓	✗	✓	✗
No demonstrations	✗	✗	✗	✗

F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

Impact of Format

- without keeping the format are overall not better than no demonstrations

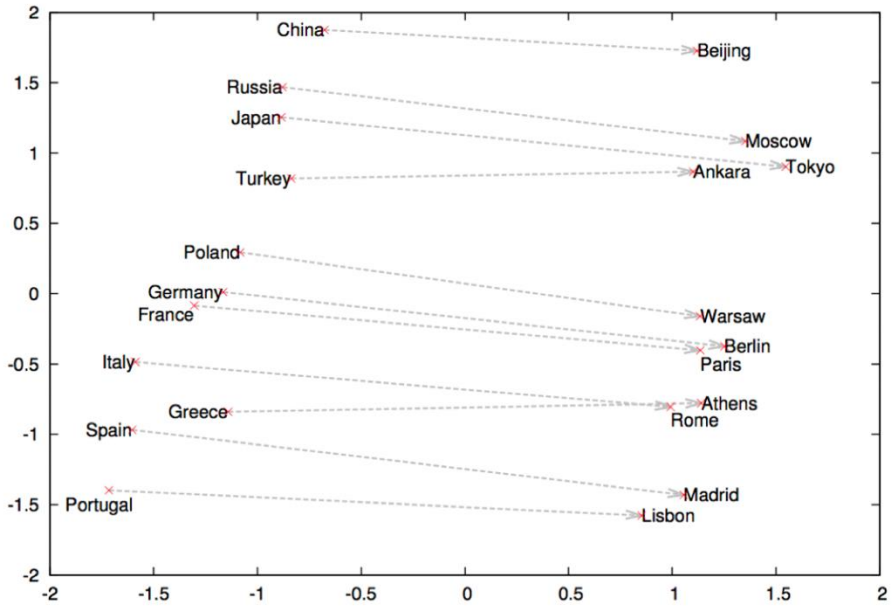


	F	*L*	*I*	*M*
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
OOD + Random labels	✓	✓	✗	✗
Random labels only	✗	✓	✗	✗
Random English words	✓	✗	✓	✗
No labels	✗	✗	✓	✗
No demonstrations	✗	✗	✗	✗

F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

Capacity of LLMs

- 单词→句子→逻辑→...



$$W(\text{"China"}) - W(\text{"Beijing"}) \approx W(\text{"Japan"}) - W(\text{"Tokyo"})$$

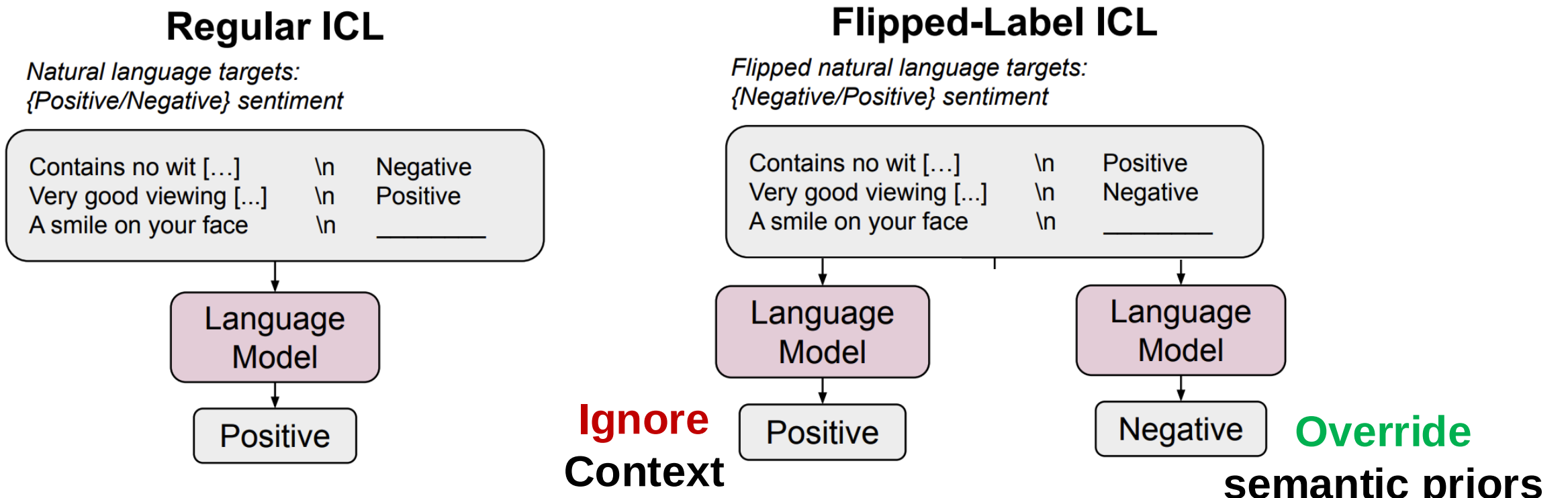
你好，我是人工智能助理ChatGLM，基于清华大学GLM-130B开发而成。我的使命是帮助人类获取信息并提供问题建议。

问题：用厚度为1厘米的玻璃做一个长42厘米，宽27厘米，深50厘米的无盖长方体容器，如果向这个容器注入30升水，水的深度是多少厘米?
回答：1升=1000立方厘米。
因为玻璃厚1厘米，所以内部的长是42-2=40厘米，内部宽是27-2=25厘米。
所以水深度为30×1000/（40×25）=30厘米。

问题：学校食堂运来19吨煤，已经烧了16吨，平均每天烧0.4吨，剩下的煤每天烧0.3吨，还要几天烧完?
回答：

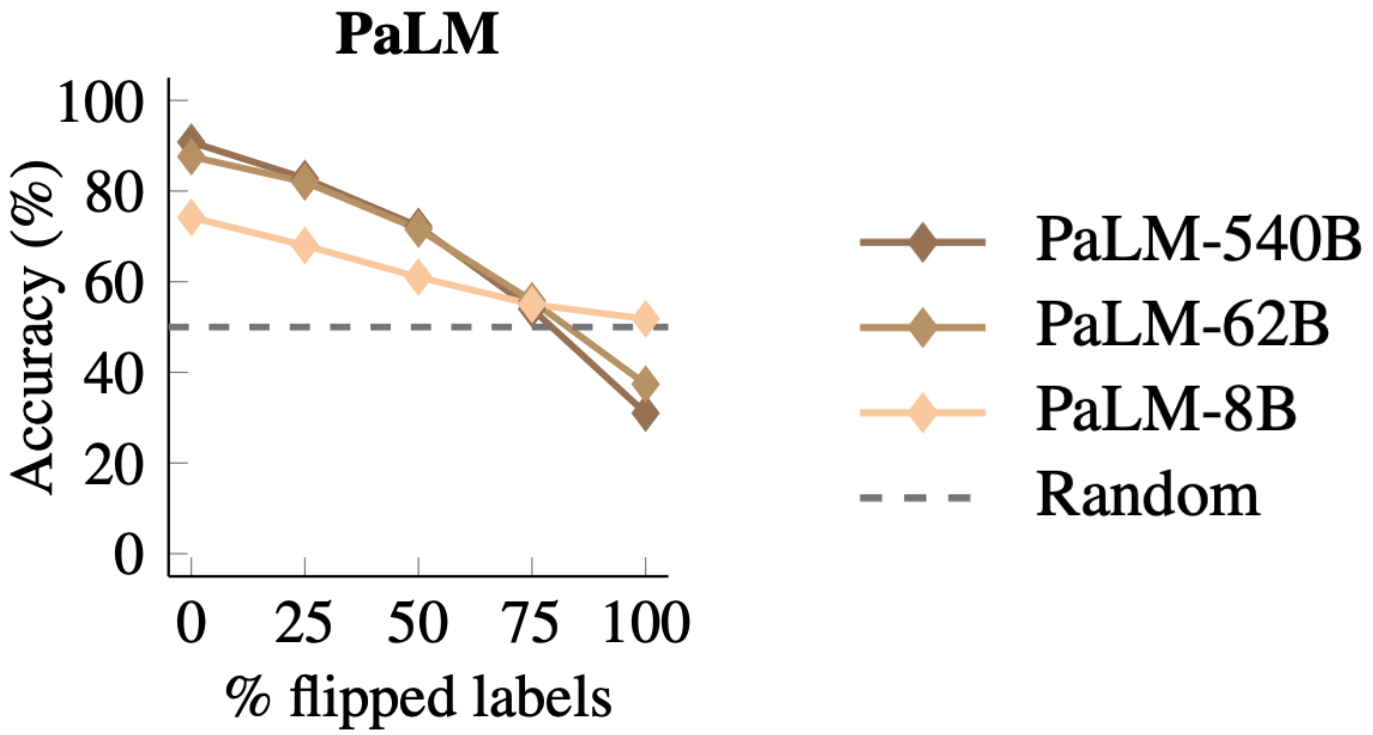
Emerging In-Context Ability in Large Models

- Large language models follow in-context instructions better
  - Can learn context mapping and ***override prior knowledge***
  - Can flip prediction with flipped/semantic-unrelated in-context labels



Emerging In-Context Ability in Large Models

- **Small models** remains relatively flat and does not below random guessing
- **Large models** experience performance drops to random guessing



How to Select Examples

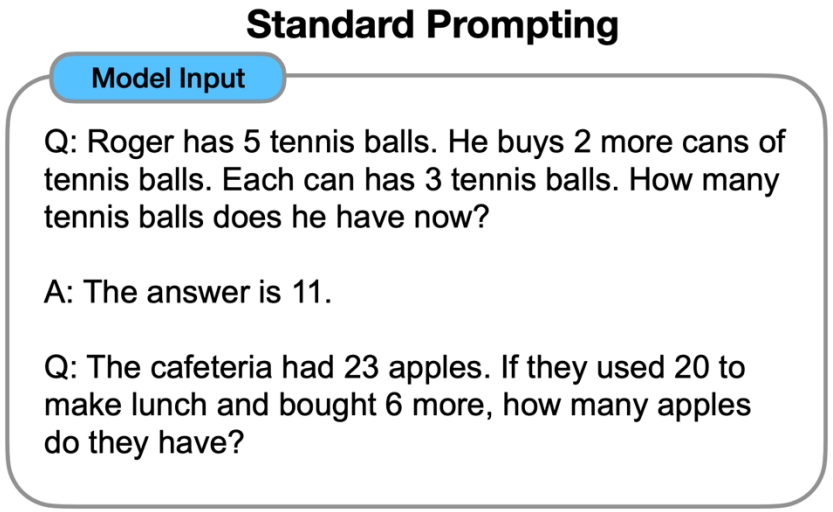
- The same example-prompt distribution is important
  - Correct label space
- More examples generally bring benefits
  - Not absolute
  - Also requiring longer context
- Correct label further benefits
- Overcome bias in in-context learning?

How to design examples

- How to better elicit the (reasoning) abilities of LLM through in-context prompting?
  - Chain of thought Prompting (CoT)
  - Self-Consistency
  - Least-to-most Prompting

Struggle of Few-shot Prompting

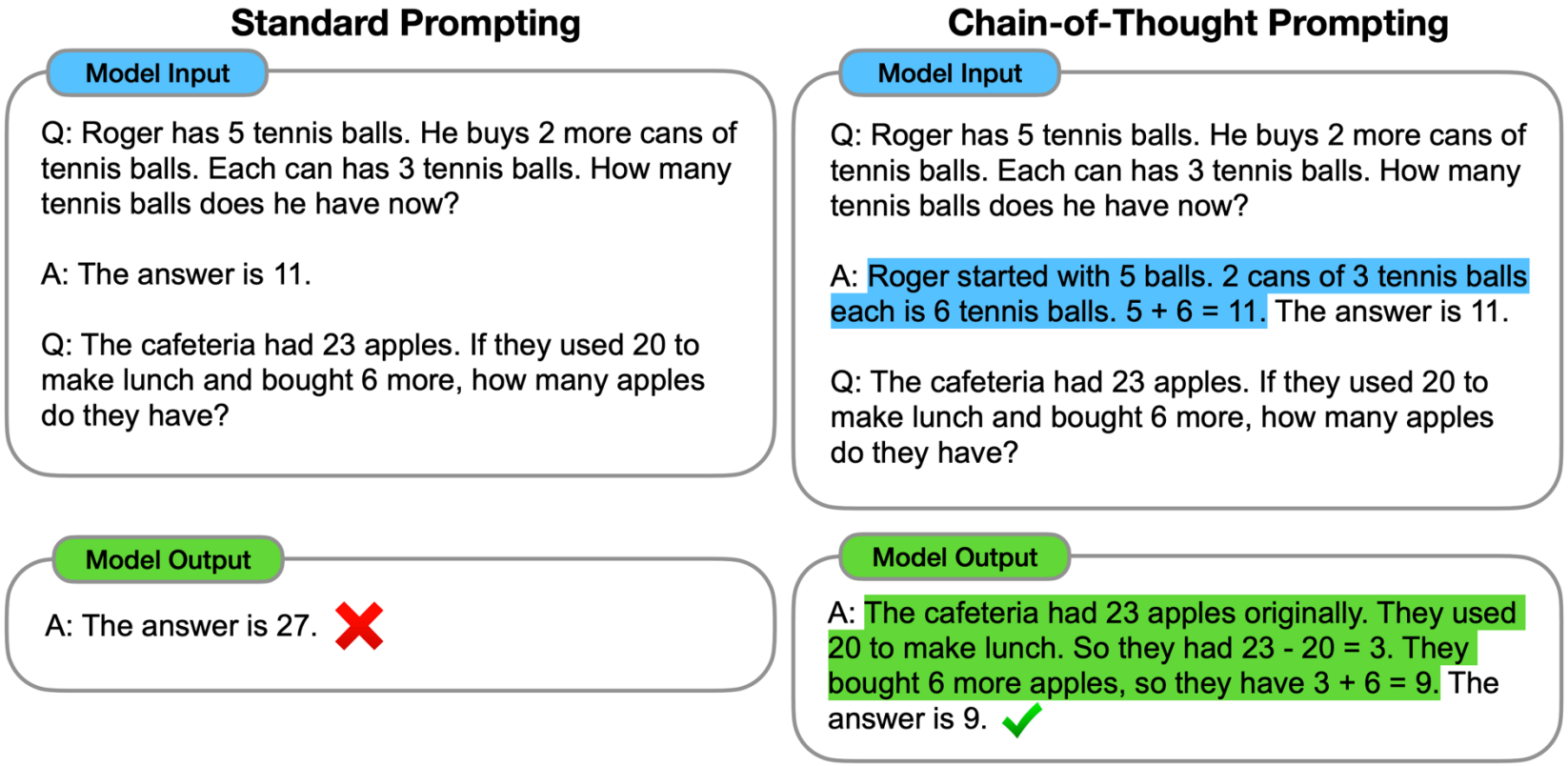
- Traditional few-shot works poorly on tasks that require reasoning abilities
  - simple input–output pairs



- Natural language rationales could benefit reasoning.
  - Gene reasoning path instead of the answer only

Chain-of-thought (CoT)

- Enable LMs to generate intermediate reasoning (*rationale*)
- Paradigm shift of task format
  - [input] → [output] ⇒ [input] → [Rationale] → [output]

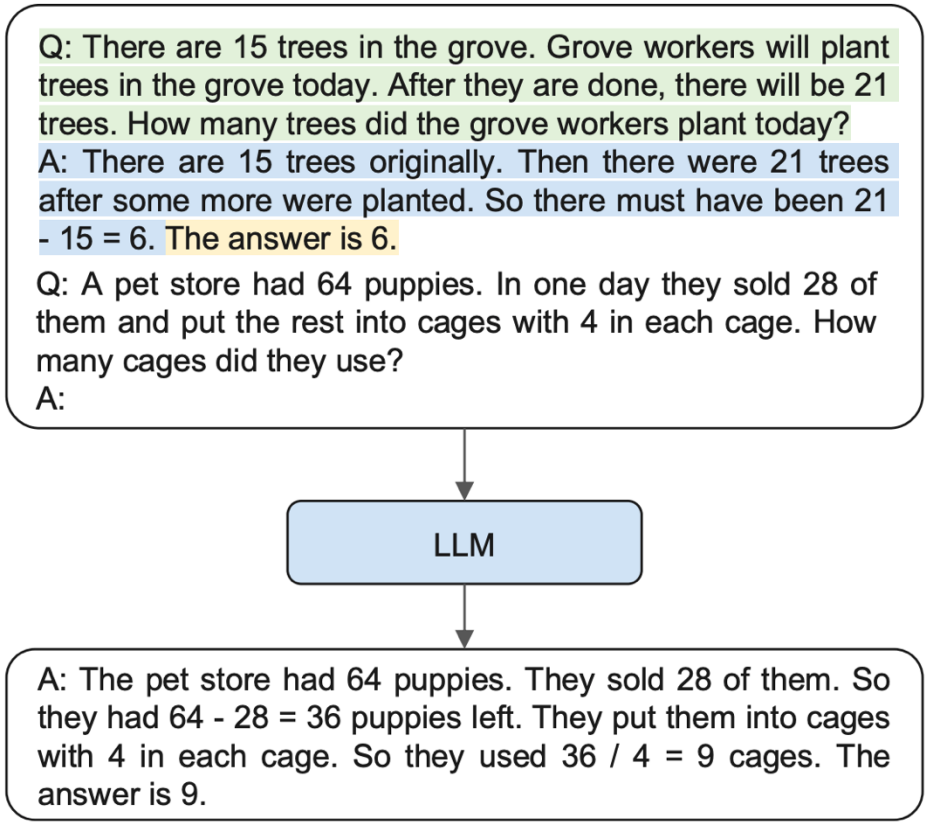


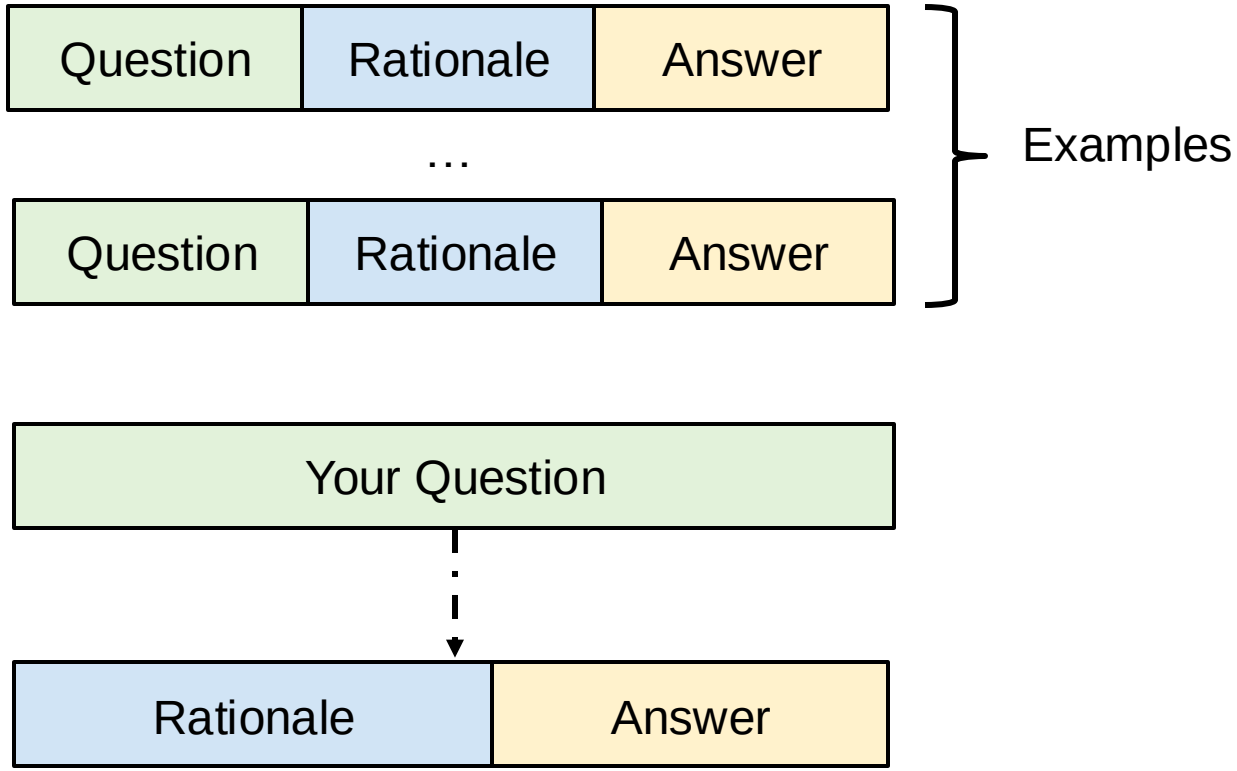
Chain-of-thought

- Enable language models to do more-complicated tasks
- Guide them with “meta-data” (i.e., reasoning process)
- Prompts are manually composed (prompt engineering helps)

Chain-of-Thought Paradigm

- Few-shot CoT
  - demonstrating step-by-step reasoning exemplars

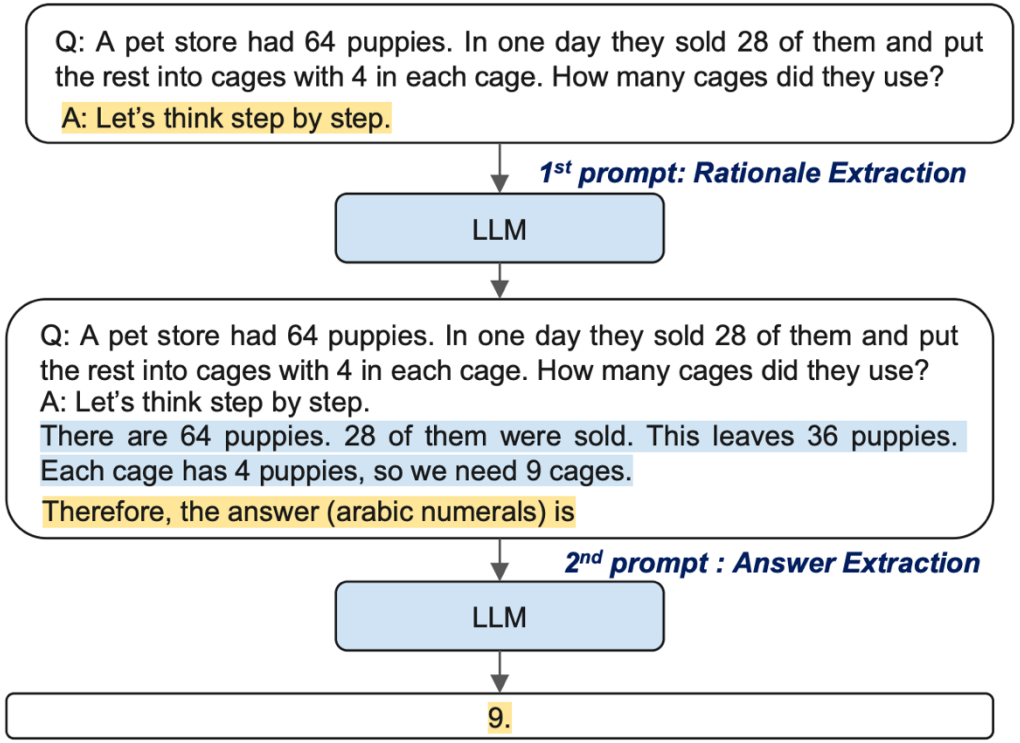


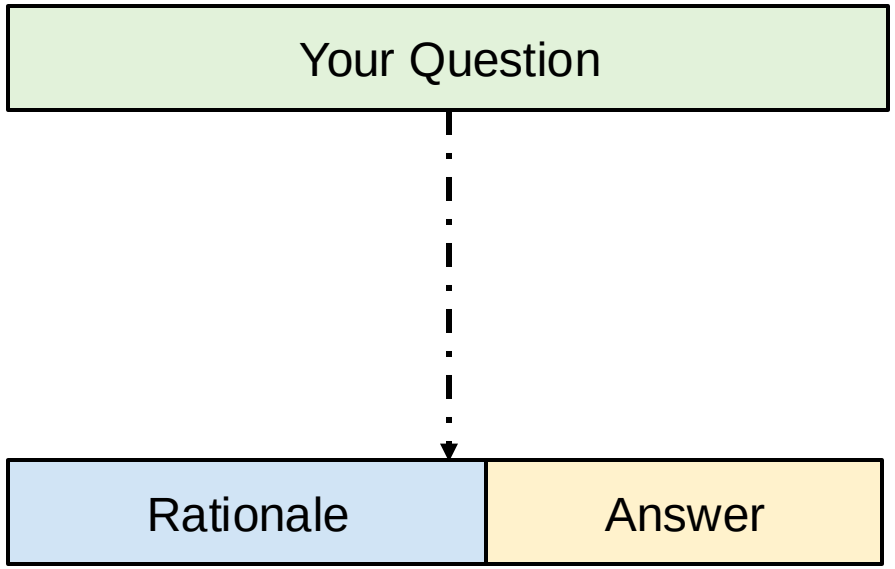


Wei et al, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Chain-of-Thought Paradigm

- Zero-shot CoT
  - Adding "***let's think step by step***" after the question can elicit the rationale of LLMs





Kojima et al, Large Language Models are Zero-Shot Reasoners

CoT elicits strong Reasoners

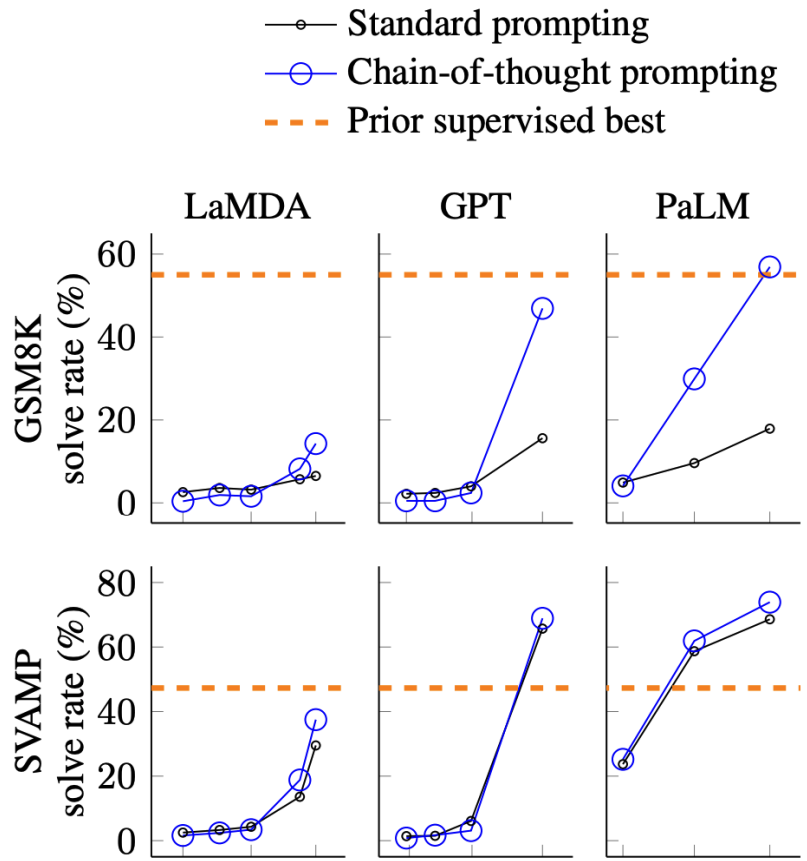
- CoT highly boosts the performance on multi-step reasoning tasks
  - Math work problems, commonsense reasoning, etc

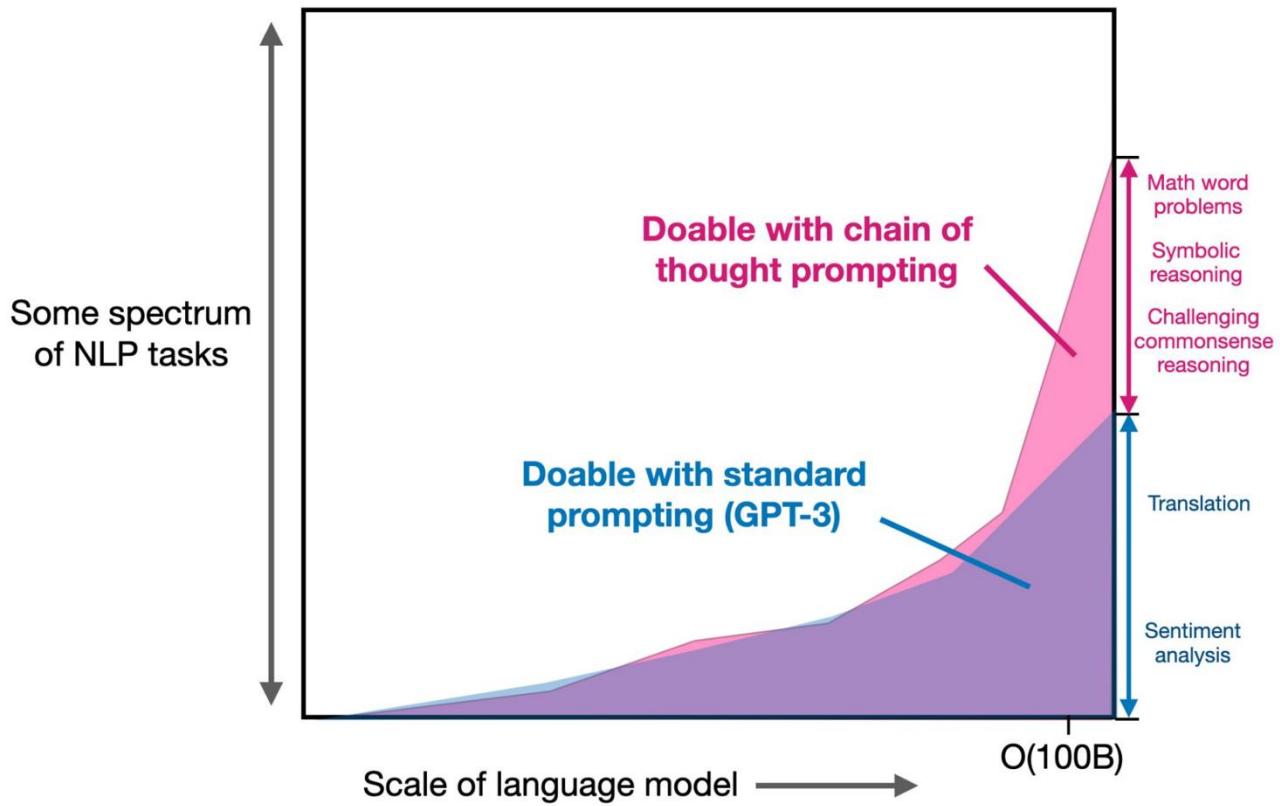
Model	MultiArith		GSM8K		AddSub		AQUA-RAT		SingleEq		SVAMP	
	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT
Zero-Shot Performance												
text-davinci-002	22.7	78.7	12.5	40.7	77.0	74.7	22.4	33.5	78.7	78.7	58.8	63.7
text-davinci-003	24.2	83.7	12.6	59.5	87.3	81.3	**28.0**	40.6	82.3	86.4	64.7	73.6
ChatGPT	**30.3**	**96.0**	**14.7**	**75.4**	**89.6**	**89.9**	23.6	**47.6**	**83.1**	**91.3**	**68.1**	**82.8**
Few-Shot Performance												
UL2	5.0	10.7	4.1	4.4	18.5	18.2	20.5	23.6	18.0	20.2	10.1	12.5
LaMDA	7.6	44.9	6.5	14.3	43.0	51.9	25.5	20.6	48.8	58.7	29.5	37.5
text-davinci-002	33.8	91.7	15.6	46.9	83.3	81.3	24.8	35.8	82.7	86.6	65.7	68.9
Codex	**44.0**	**96.2**	**19.7**	**63.1**	90.9	90.9	**29.5**	**45.3**	**86.8**	**93.1**	**69.9**	76.4
PaLM	42.2	94.7	17.9	56.9	**93.9**	**91.9**	25.2	35.8	86.5	92.3	69.4	**79.0**

Qin et al, Is ChatGPT a General-Purpose Natural Language Processing Task Solver?

Scaling Property

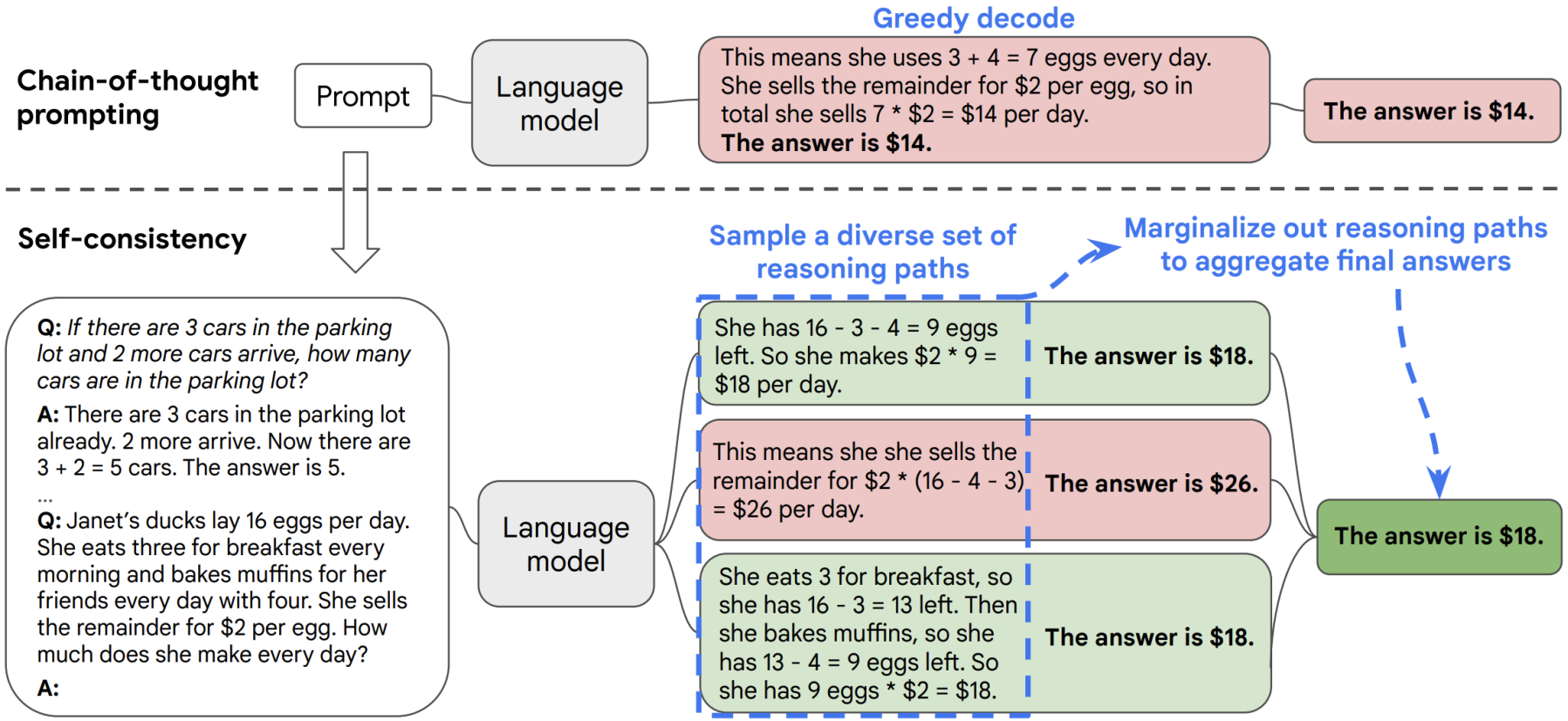
- Chain-of-thought shows better scaling properties in handling complex problems





CoT with Self-Consistency

- Further boost CoT by selecting the most consistent result from a diverse set of reasoning paths
  - A complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer.



CoT with Self-Consistency

1. Elicit different reasoning path with different in-context examples
2. Get the final answer by majority voting

$$P(\mathbf{r}_i, \mathbf{a}_i \mid \text{prompt}, \text{question}) = \exp \frac{1}{K} \sum_{k=1}^{K} \log P(t_k | \text{prompt}, \text{question}, t_1, \ldots, t_{k-1})$$

$\boldsymbol{r_i}$: reasoning path; $\boldsymbol{a_i}$: answer

	GSM8K	MultiArith	AQuA	SVAMP	CSQA	ARC-c
Greedy decode	56.5	94.7	35.8	79.0	79.0	85.2
Weighted avg (unnormalized)	56.3 ± 0.0	90.5 ± 0.0	35.8 ± 0.0	73.0 ± 0.0	74.8 ± 0.0	82.3 ± 0.0
Weighted avg (normalized)	22.1 ± 0.0	59.7 ± 0.0	15.7 ± 0.0	40.5 ± 0.0	52.1 ± 0.0	51.7 ± 0.0
Weighted sum (unnormalized)	59.9 ± 0.0	92.2 ± 0.0	38.2 ± 0.0	76.2 ± 0.0	76.2 ± 0.0	83.5 ± 0.0
Weighted sum (normalized)	74.1 ± 0.0	99.3 ± 0.0	48.0 ± 0.0	86.8 ± 0.0	80.7 ± 0.0	88.7 ± 0.0
Unweighted sum (majority vote)	74.4 ± 0.1	99.3 ± 0.0	48.3 ± 0.5	86.6 ± 0.1	80.7 ± 0.1	88.7 ± 0.1

Why does In-Context Learning work?

In-Context Learning by Gradient Descent

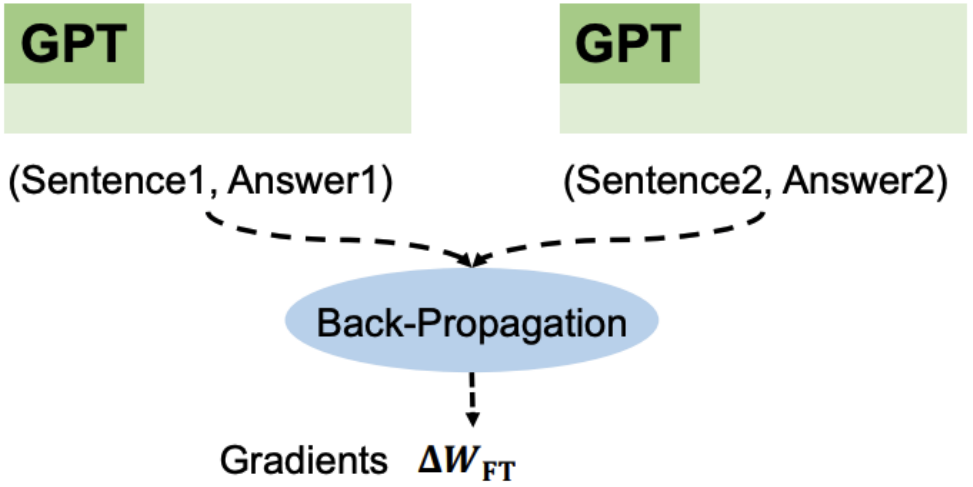
Why does In-Context Learning work?

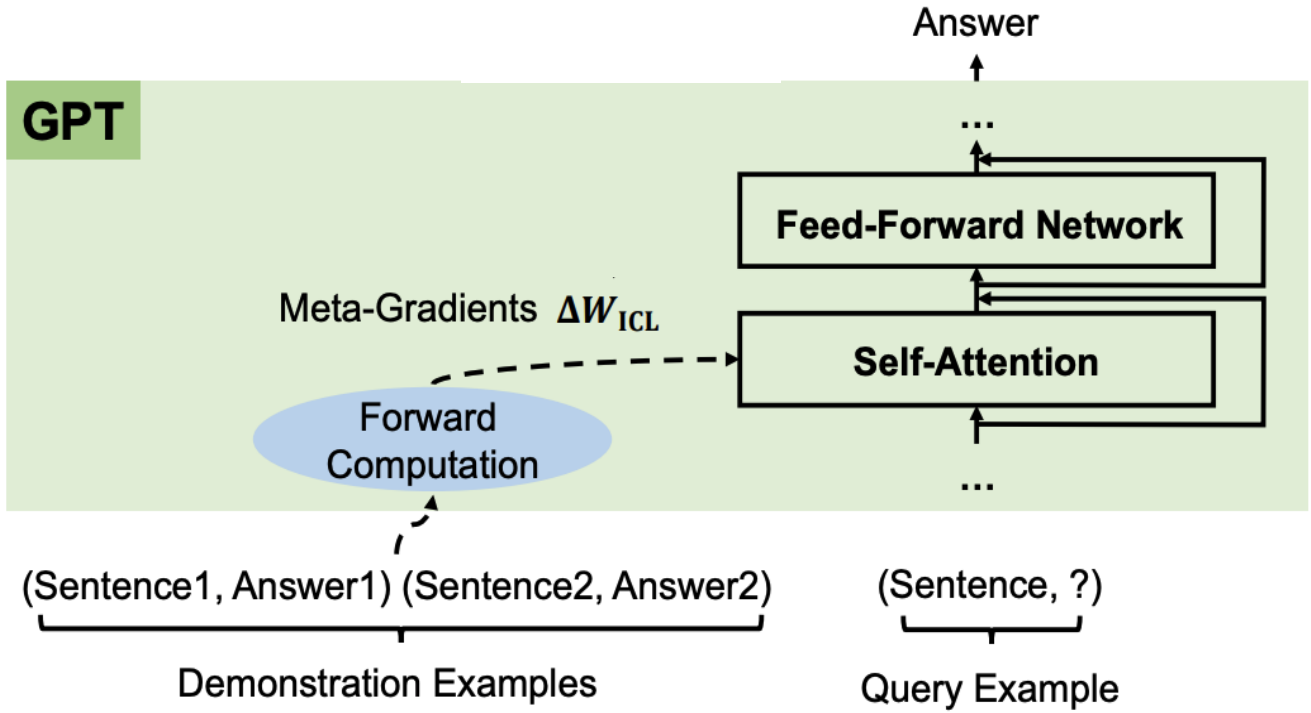
- Till now, the mechanisms of in-context learning in Transformers are still not well understood and remain mostly an intuition.
  - associative memory (Ramsauer et al., 2020)
  - copying mechanism termed *induction heads* (Olsson et al., 2022)
- ***New observations***:
  - Training Transformers on auto-regressive objectives is closely related to gradient-based meta-learning
  - Trained Transformers become meta-optimizers i.e. learn models by gradient descent in their forward pass

- Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." *International Conference on Machine Learning*. PMLR, 2023.
- Ramsauer, Hubert, et al. "Hopfield networks is all you need." *arXiv preprint arXiv:2008.02217* (2020).
- Olsson, Catherine, et al. "In-context learning and induction heads." *arXiv preprint arXiv:2209.11895* (2022).

Why In-Context Learning Works

- Rethnking the connection between in-context learning and finetuning





Finetuning:
Use examples to ***update*** parameters

In-Context Learning:
Use examples to ***activate*** parameters

Why In-Context Learning Works ≈ Gradient Descent

In-contexts as implicit gradient descent for attention

- *Attention decomposition perspective*

X' =demonstration tokens
X =query tokens

$$\mathcal{F}_{ICL}(\mathbf{q}) = \mathrm{Attn}(V, K, \mathbf{q})$$

$$= W_V[X'; X]\mathrm{softmax}\left(\frac{(W_K[X'; X]^\top \mathbf{q})}{\sqrt{d}}\right)$$

Ignore softmax

$$\approx W_V[X'; X](W_K[X'; X]^\top)\mathbf{q}$$

$$= W_V X(W_K X)^\top \mathbf{q} + W_V X'(W_K X')^\top \mathbf{q}$$

Context

Zero-shot

$$= (W_{ZSL} + W_{ICL})q$$

$$\mathcal{F}_{FT}(\mathbf{q}) = (W_{ZSL} + \Delta W_{FT})\mathbf{q}$$

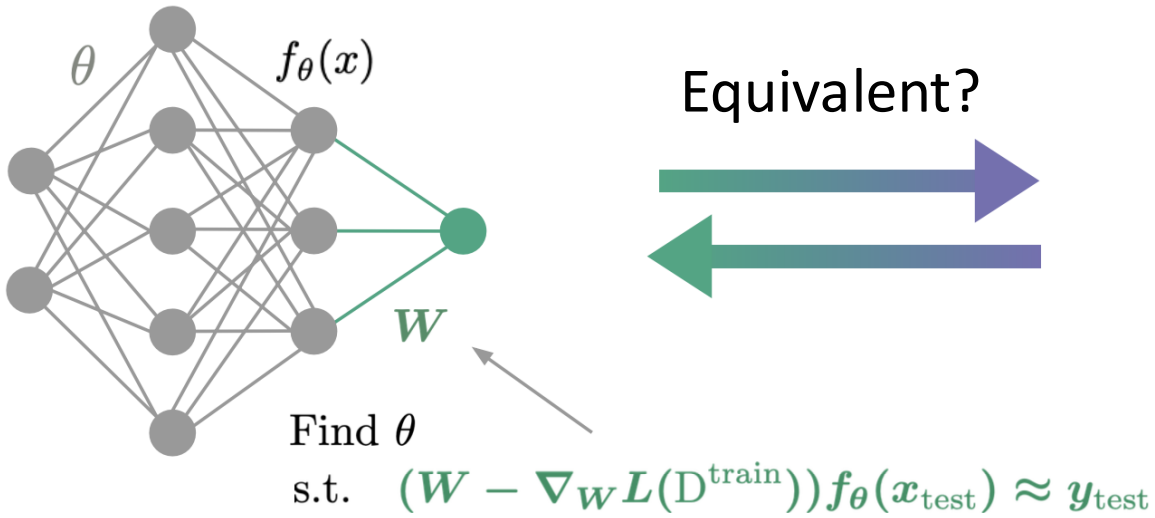
dual form of the Transformer
attention = A linear layer
optimized by gradient descent

Dai et al, Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers

Transformers Learn In-Context by Gradient Descent

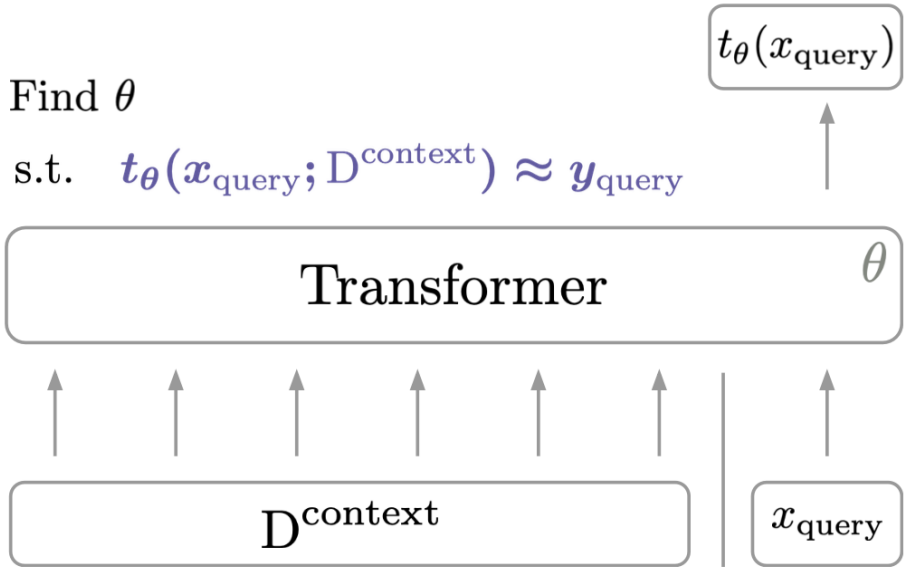
- *Data transformation perspective*

Learning an output layer
by gradient descent



Equivalent?

Adjusts query prediction given in-context



Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." *International Conference on Machine Learning*. PMLR, 2023.

Transformers Learn In-Context by Gradient Descent

- Gradient descent formulated as data transformations
- Weights updating as data transformation.
  - Linear regression as a simple example

$$L(W) = \frac{1}{2N}\sum_{i=1}^{N}\|Wx_i - y_i\|^2.$$

$$\Delta W = -\eta\nabla_W L(W) = -\frac{\eta}{N}\sum_{i=1}^{N}(Wx_i - y_i)x_i^T.$$

One step of gradient descent on L with learning rate η yields the weight change

$$L(W+\Delta W) = \frac{1}{2N}\sum_{i=1}^{N}\|(W+\boxed{\Delta W})x_i - y_i\|^2$$

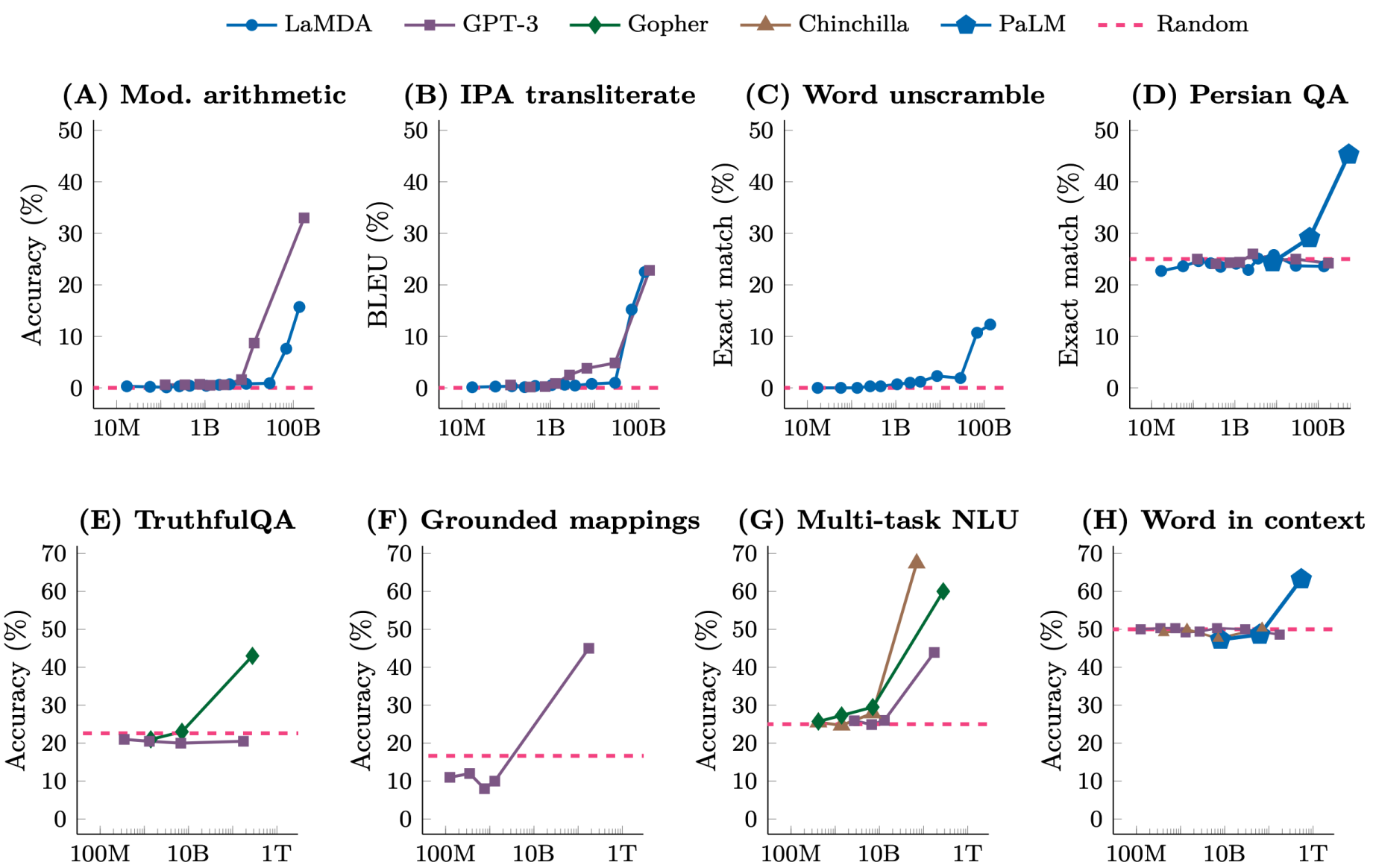
$$= \frac{1}{2N}\sum_{i=1}^{N}\|Wx_i - (y_i - \boxed{\Delta y_i})\|^2$$

Discussion

- Gradient-based optimization ⇔ Attention-based in-context learning
- Finetuning ⇔ In-context learning
- **Viewed through the lens of meta-learning**: learning Transformer weights corresponds to the outer-loop which then enables the forward pass to transform tokens by gradient-based optimization.
- **Meta-learned task-shared learning rates**. When training self-attention parameters θ across a family of in-context learning tasks, the learning rate can be implicitly (meta-)learned such that an optimal loss reduction (averaged over tasks) is achieved.

Understanding Emergent Abilities

Emergent Abilities (Wei, et al., 2022)



J Wei, et al. Emergent Abilities of Large Language Models. arXiv: 2206.07682

Reasoning

- Multi-step reasoning, Instruction following, Program execution, Model calibration.

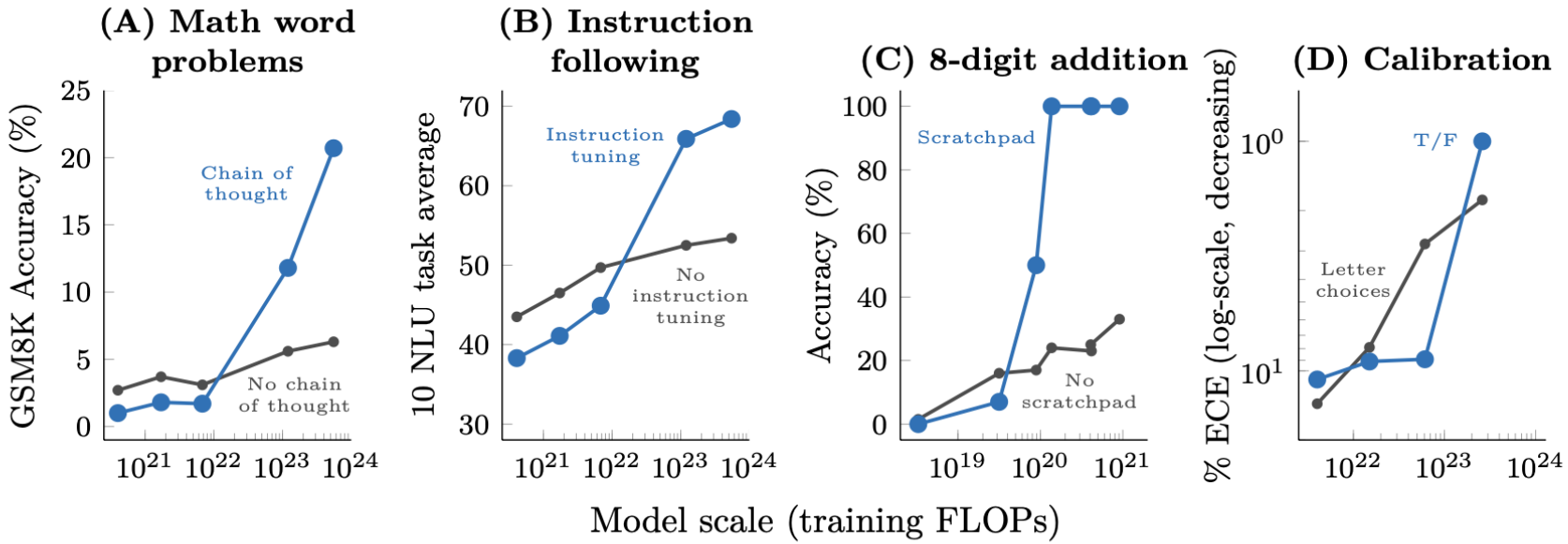


Figure 3: Specialized prompting or finetuning methods can be emergent in that they do not have a positive effect until a certain model scale. A: Wei et al. (2022b). B: Wei et al. (2022a). C: Nye et al. (2021). D: Kadavath et al. (2022). An analogous figure with number of parameters on the x-axis instead of training FLOPs is given in Figure 12. The model shown in A-C is LaMDA (Thoppilan et al., 2022), and the model shown in D is from Anthropic.

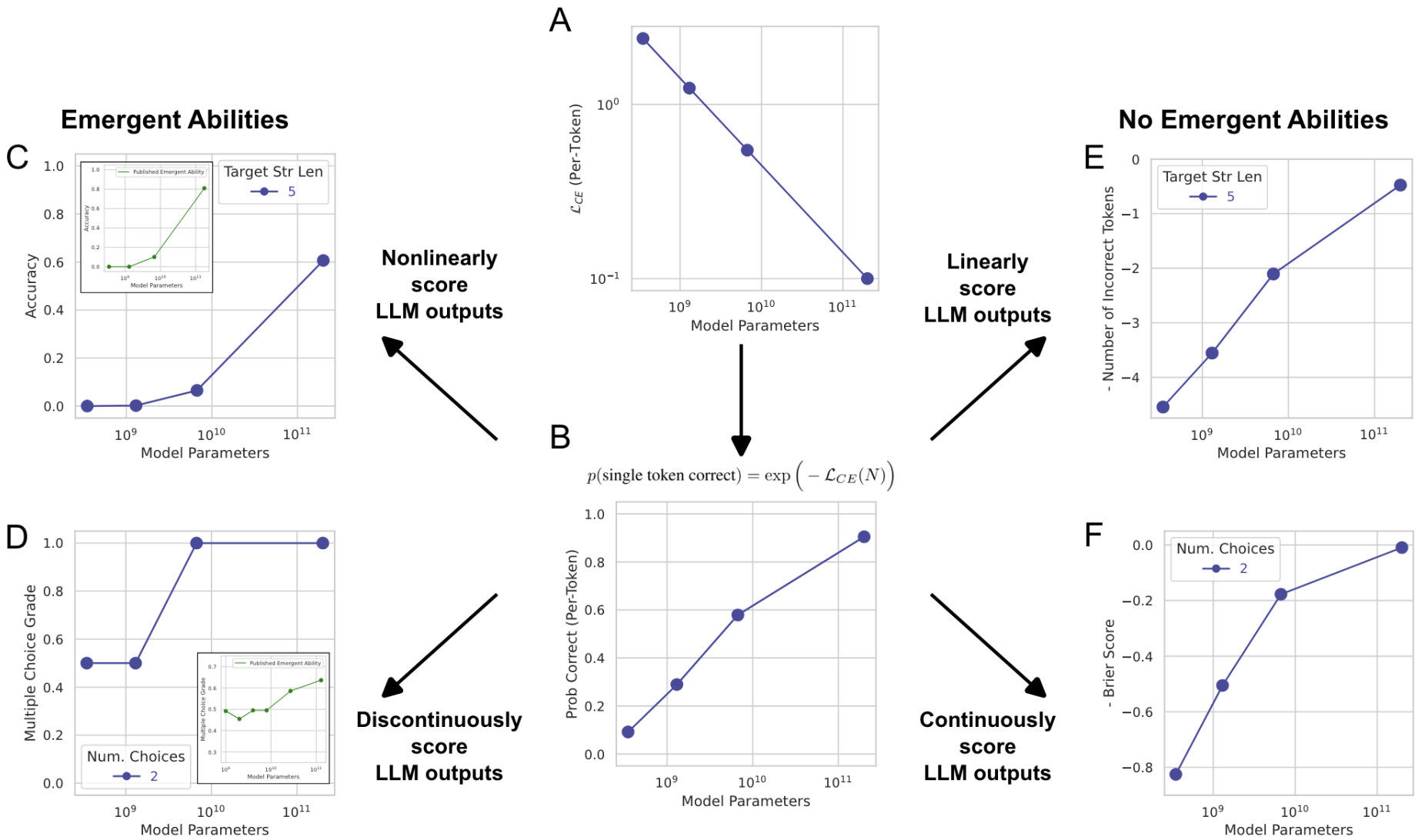
Are Emergent Abilities of Large Language Models a Mirage? (Schaeffer, et al., NeurIPS'23 Best Paper)

- **Sharpness**, transitioning seemingly instantaneously from not present to present
- **Unpredictability**, transitioning at seemingly unforeseeable model scales

- Emergent Abilities Questions:
  - What controls **which** abilities will emerge?
  - What controls **when** abilities will emerge?

Rylan Schaeffer, Brando Miranda, Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? NeurIPS'23.

Are Emergent Abilities of Large Language Models a Mirage? (Schaeffer, et al., NeurIPS'23 Best Paper)



Rylan Schaeffer, Brando Miranda, Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? NeurIPS'23.

Are Emergent Abilities of Large Language Models a Mirage? (Schaeffer, et al., NeurIPS'23 Best Paper)

- Assume ***per-token cross entropy*** falls as a power law

$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha}$$

where N is the number of parameters,

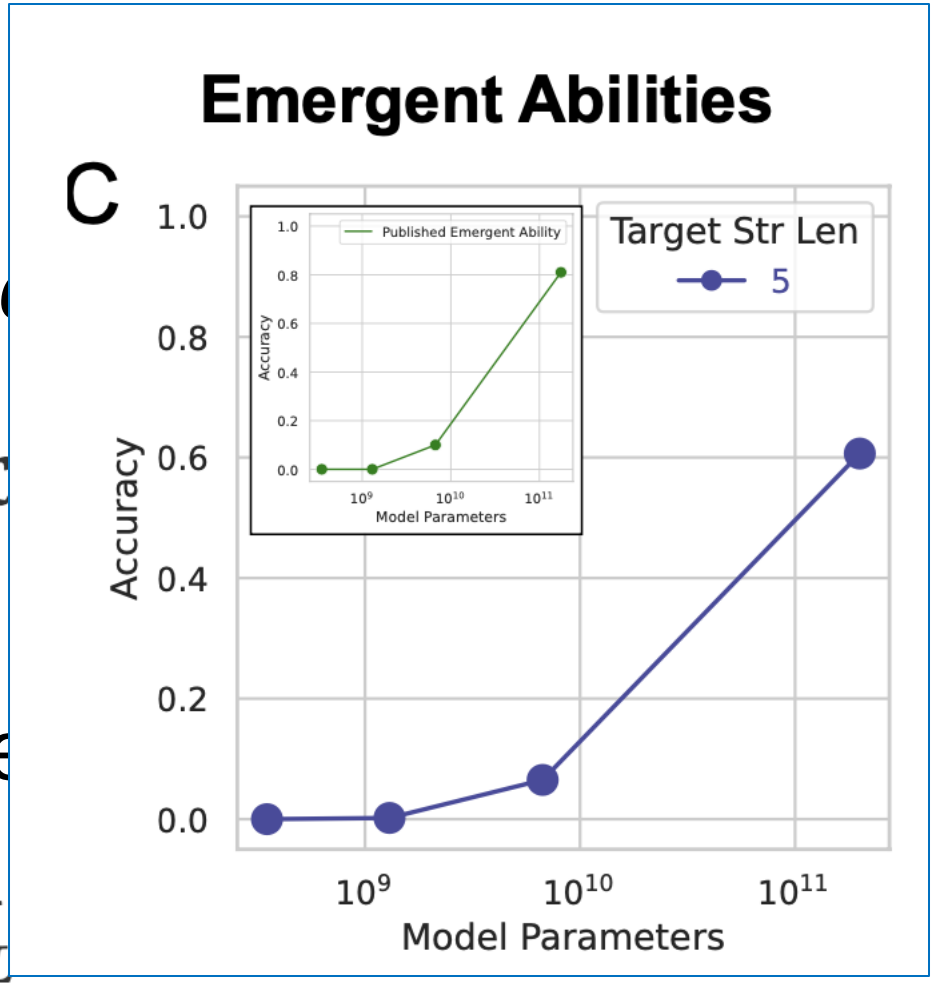
$$\mathcal{L}_{CE}(N) \overset{\text{def}}{=} -\sum_{v \in V} p(v) \log \hat{p}_N(v) \quad \Rightarrow \quad \mathcal{L}$$

$$p(v^*) = 1$$

If a metric that requires selecting L toke

$$\text{Accuracy}(N) \approx p_N(\text{single token}$$

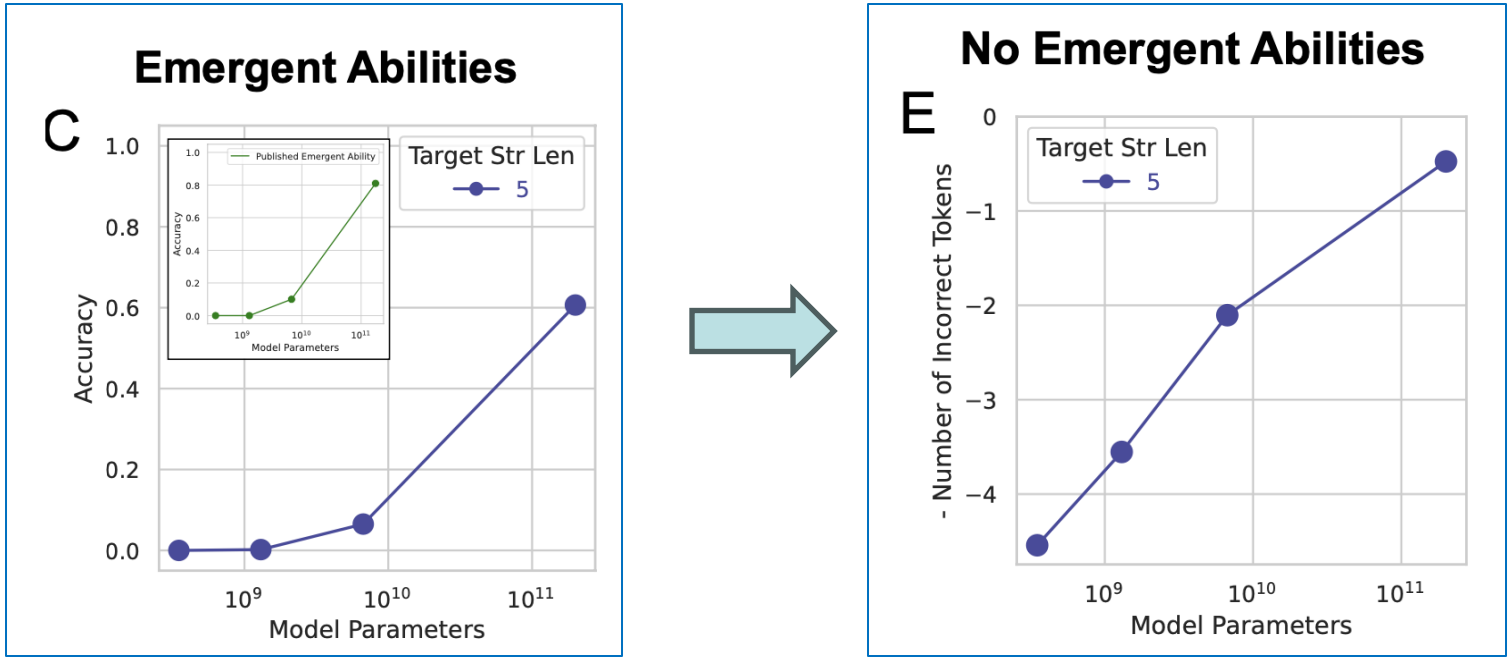
$$= \exp\left(-(N/c)^{\alpha}\right)^{L}$$



Are Emergent Abilities of Large Language Models a Mirage? (Schaeffer, et al., NeurIPS'23 Best Paper)

- But if we use an approximately linear metric, e.g., Token Edit Distance

$$\text{Token Edit Distance}(N) \approx L(1 - p_N(\text{single token correct}))$$
$$= L\left(1 - \exp\left(-(N/c)^{\alpha}\right)\right)$$



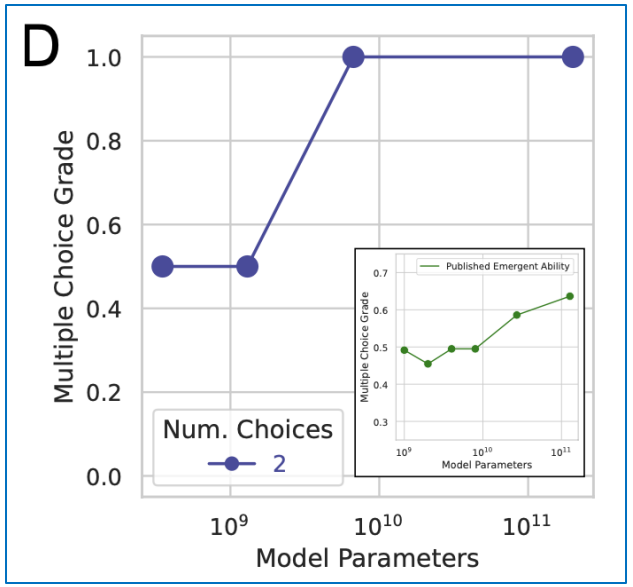
Are Emergent Abilities of Large Language Models a Mirage? (Schaeffer, et al., NeurIPS'23 Best Paper)

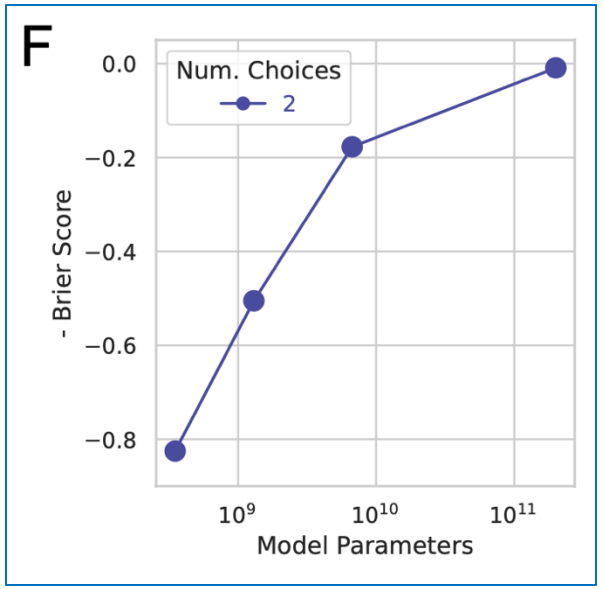
Multiple Choice Grade:

$$\text{Multiple Choice Grade} \overset{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on co} \\ 0 & \text{otherwise} \end{cases}$$

Brier Score:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$





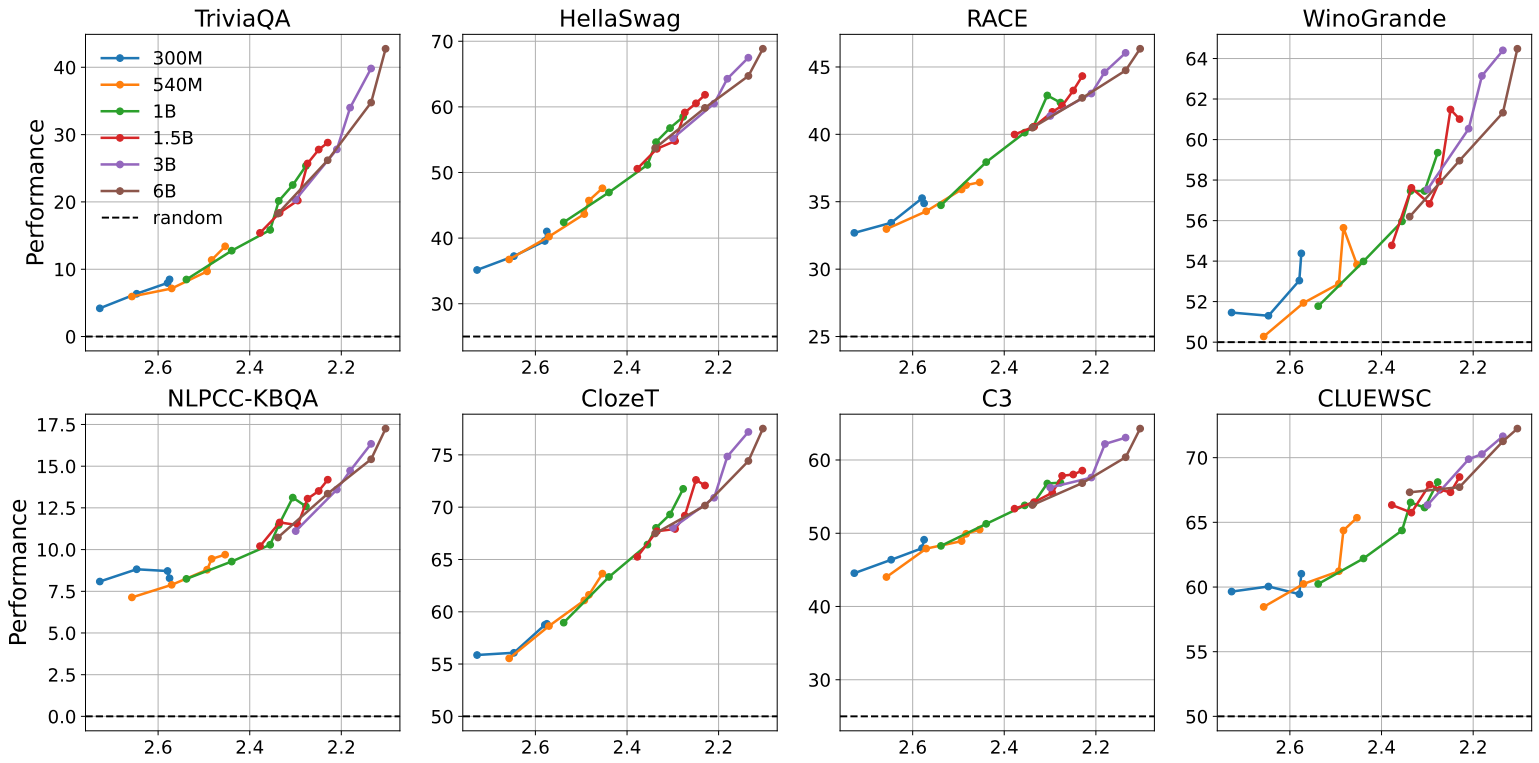
Questions?

- Before, people believe that emergent abilities are exclusive to LLMs
- But now
  - smaller models can also exhibit high performance on emergent abilities
  - continuous metrics “seems” no emergent abilities

Do you believe LLMs having “emergent ability”?

Understanding Emergent Abilities of Language Models from the Loss Perspective? (Du, et al., NeurIPS'24)

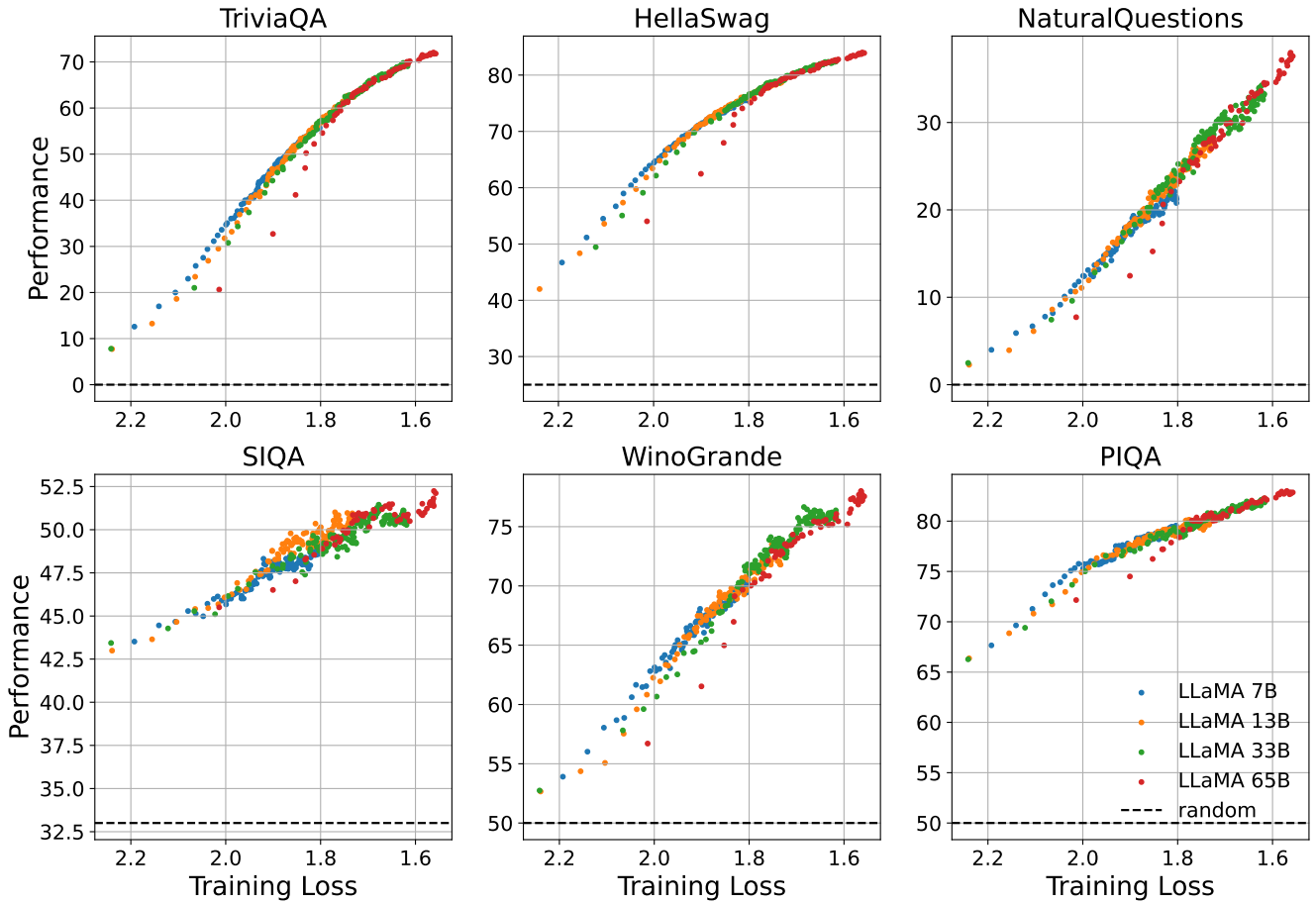
Relationship between Pre-training Loss and Downstream Performance



We train a series of language models with different sizes and training tokens. We find that the data points of different sizes and training tokens largely fall on the same trending curves. In other words, the LMs with the same pre-training loss regardless of token count and model size exhibit the same performance on different languages, tasks and prompting formats.

Understanding Emergent Abilities of Language Models from the Loss Perspective? (Du, et al., NeurIPS'24)

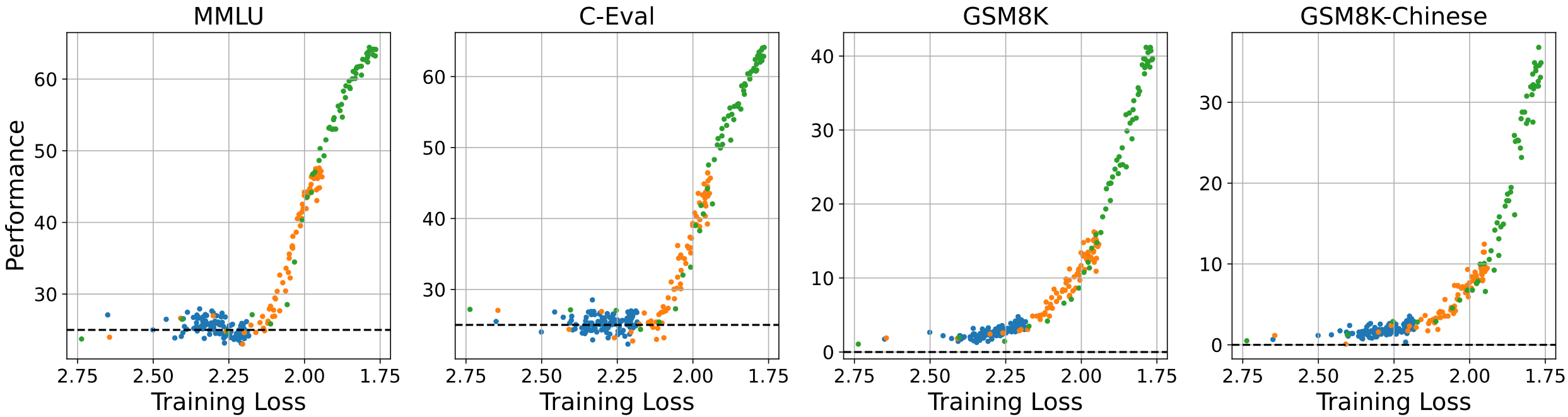
Observation from LLaMA's Paper



We extract LLaMA's pre-training loss and corresponding performance on six question answering and commonsense reasoning tasks from the figures in the original paper. Most data points from the LLaMA models fall on the same upwards trend, except those at the early stage of LLaMA-65B.

Understanding Emergent Abilities of Language Models from the Loss Perspective? (Du, et al., NeurIPS'24)

Emergent Curves from the from the Loss Perspective



On MMLU, C-Eval, GSM8K, and GSM8K-Chinese, all models of three sizes perform at the random level until the pre-training loss decreases to about 2.2, after which the performance gradually climbs as the loss increases.

Defining Emergent Abilities from the Loss Perspective

The normalized performance on an emergent ability as a function of the pre-training loss L is:

$$\begin{cases} f(L) & \text{if } L < \eta \\ 0 & \text{otherwise} \end{cases}$$

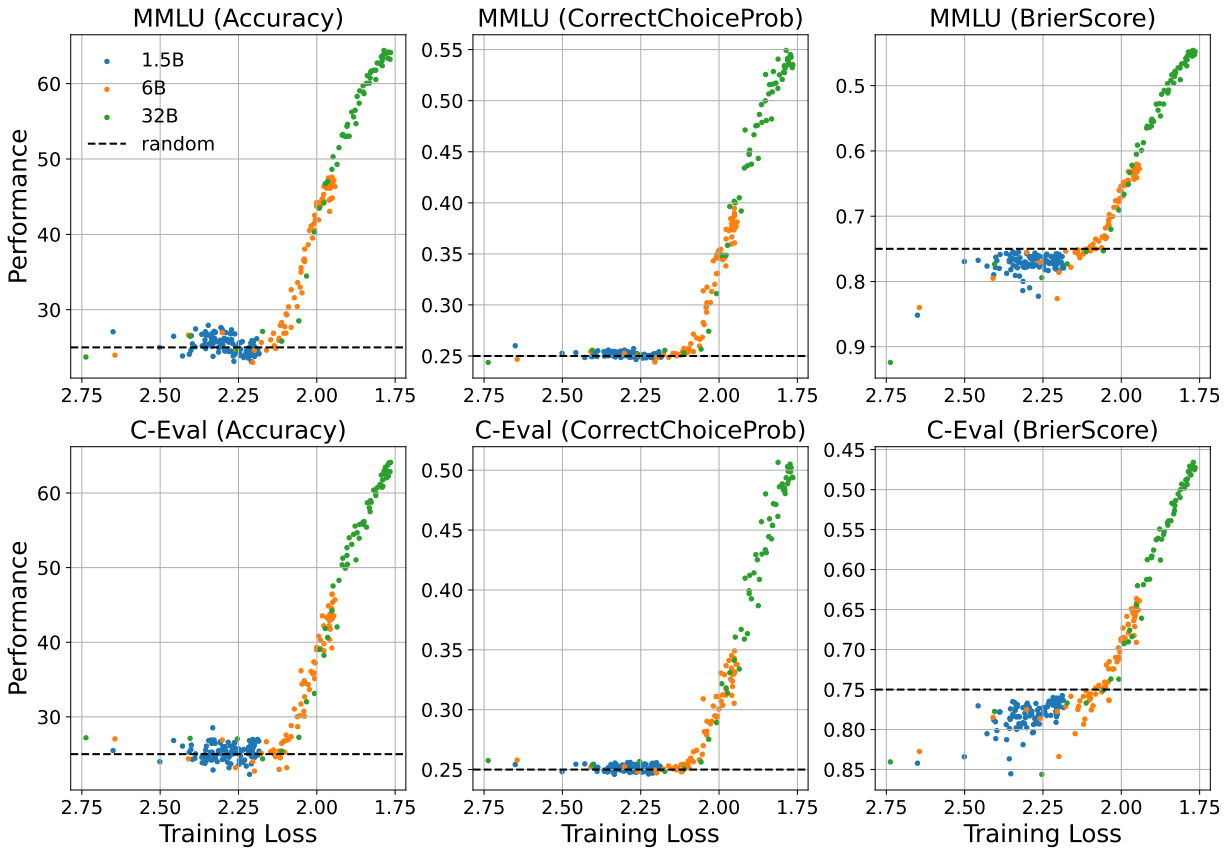
Combined with the model scaling law, we can get the normalized performance as a function of the model size N

$$\begin{cases} f\left(L_\infty + \left(\frac{N_0}{N}\right)^{\alpha_N}\right) & \text{if } N \geq N_0(\eta - L_\infty)^{-\frac{1}{\alpha^N}} \\ 0 & \text{otherwise} \end{cases}$$

From this equation, we can explain the observed emergent abilities with model sizes.

Understanding Emergent Abilities of Language Models from the Loss Perspective? (Du, et al., NeurIPS'24)

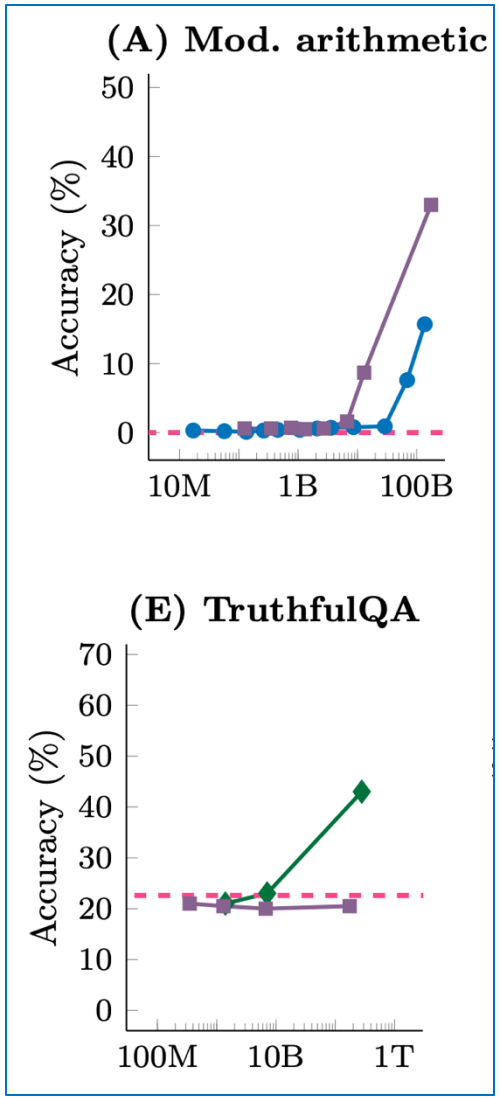
Influence of Different Metrics



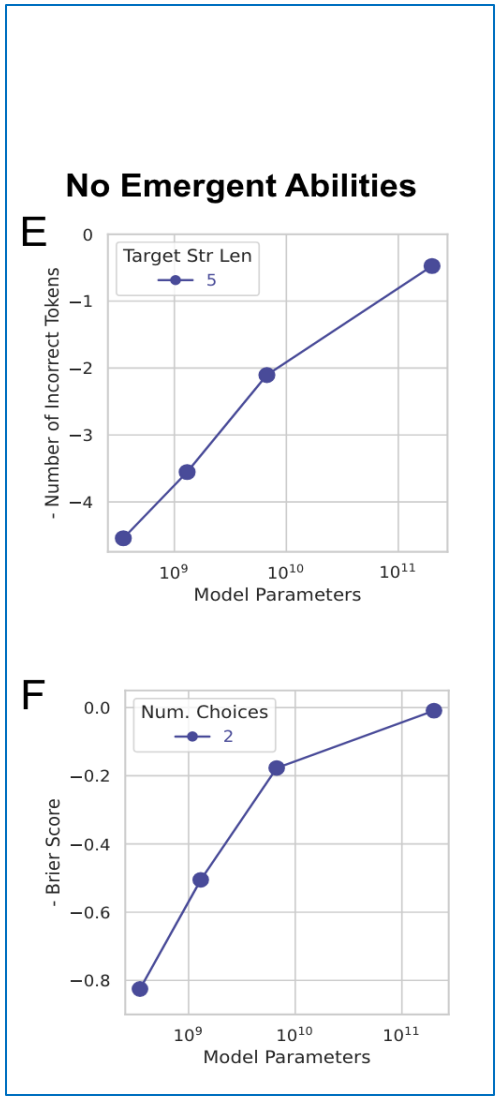
Both discontinuous and continuous metrics (accuracy, correct choice probability, and Brier Score) — show emergent performance improvements (value increase for the first two and decrease for the third) when the pre-training loss drops below a certain threshold.

Emergent Abilities

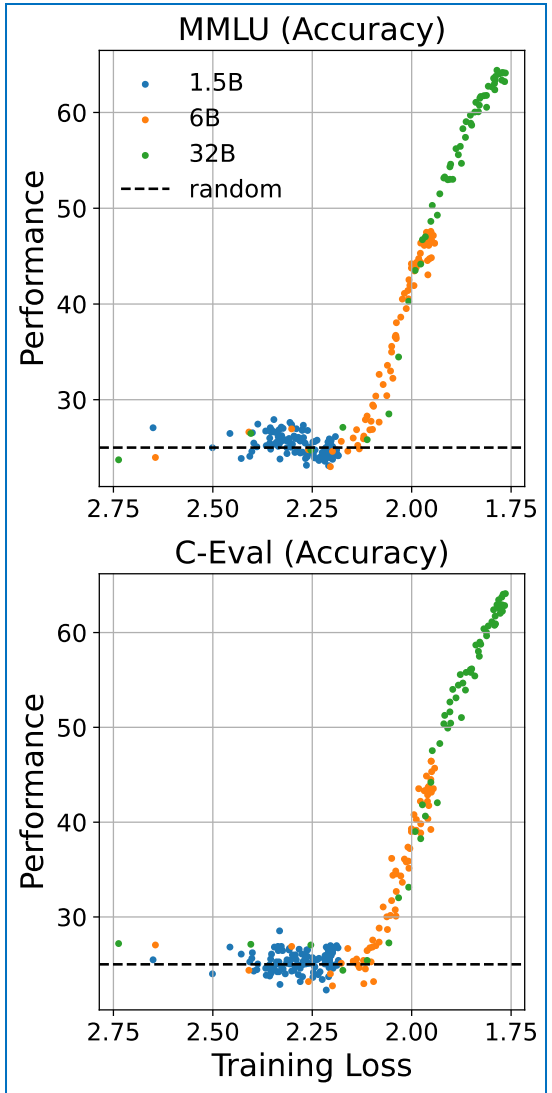
J Wei, et al. Emergent Abilities of Large Language Models. TMLR, 2022.



Schaeffer, et al., Are Emergent Abilities of Large Language Models a Mirage? NeurIPS'23



Du, et al., Understanding Emergent Abilities of Language Models from the Loss Perspective? NeurIPS'24



Summary

- In-context Learning
  - What is in-context learning?
  - Context Construction
  - Theoretical understanding of In-Context Learning
- Understanding Emergent Abilities
  - J Wei, et al. Emergent Abilities of Large Language Models. TMLR, 2022.
  - Schaeffer, et al., Are Emergent Abilities of Large Language Models a Mirage? NeurIPS'23
  - Du, et al., Understanding Emergent Abilities of Language Models from the Loss Perspective? NeurIPS'24

In-Context Learning & Emergent Abilities

Jie Tang, KEG, Tsinghua University

http://keg.cs.tsinghua.edu.cn/jietang

https://github.com/THUDM