



10-423/10-623 Generative AI

Machine Learning Department
School of Computer Science
Carnegie Mellon University

VAEs and Variational Inference & In

Matt Gormley & Pat Virtue

Lecture 8

Feb. 10, 2025

Reminders

- **Quiz 2:**
 - In-class: Mon, Feb 17
 - Lectures 5-8 + 9 (but not In-context Learning)
- **Homework 2: Generative Models of Images**
 - Due: Sat, Feb 22 at 11:59pm

Q & A

Q: I just asked a question in OH and now my TA is crying quietly what did I do wrong?

A: You've just committed the worst of crimes: asking a question that was directly answered in a recitation.

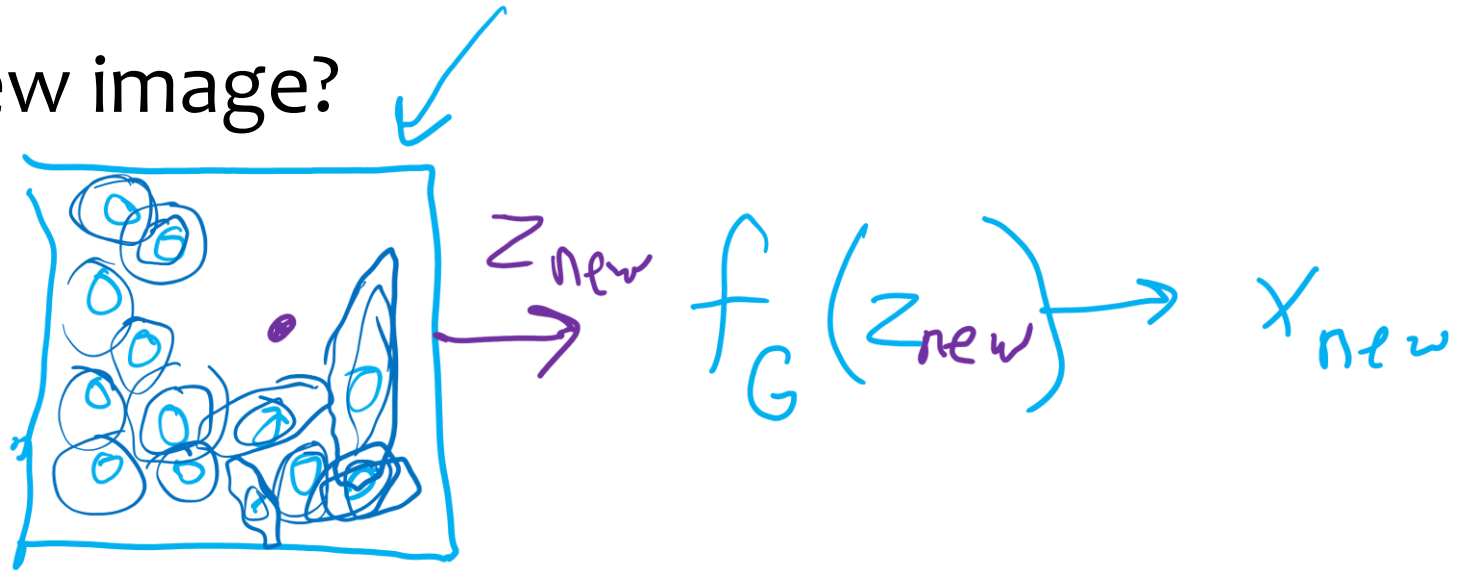
The TA you asked spent hours carefully writing careful recitation notes and solutions, practicing their recitation, responding to criticism / changes from instructors, etc.

To increase OH efficiency, please review the HW recitation before asking HW questions in OHs.

VARIATIONAL AUTOENCODERS

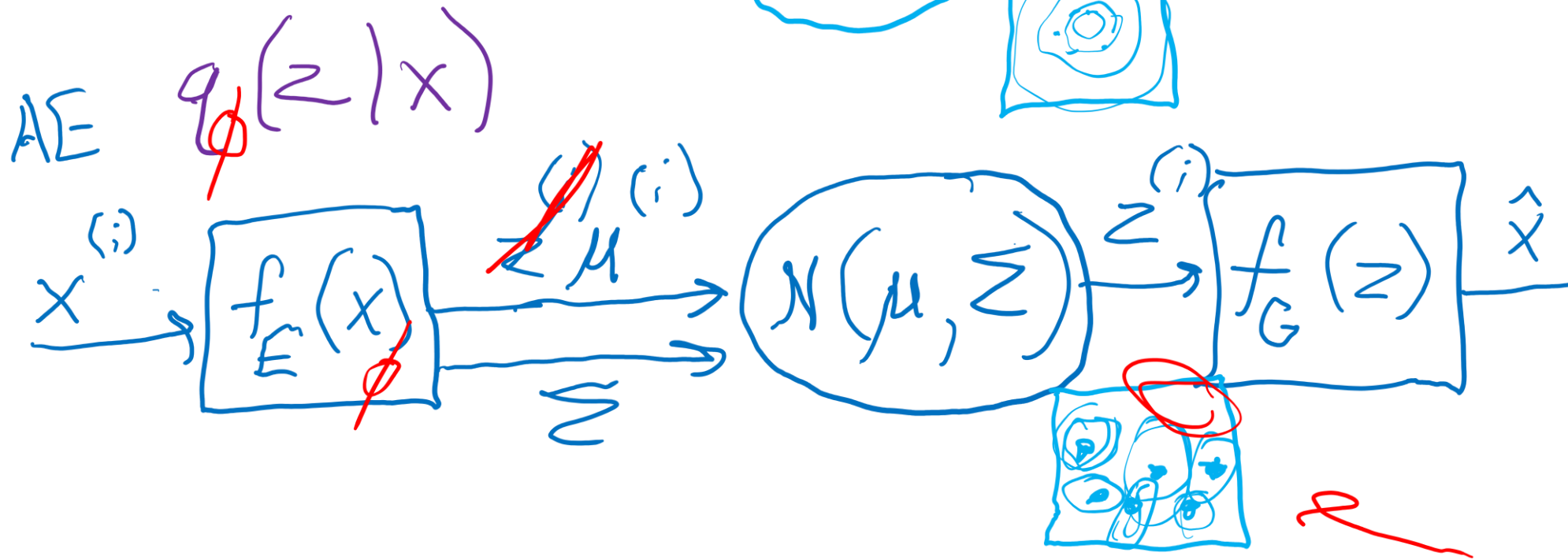
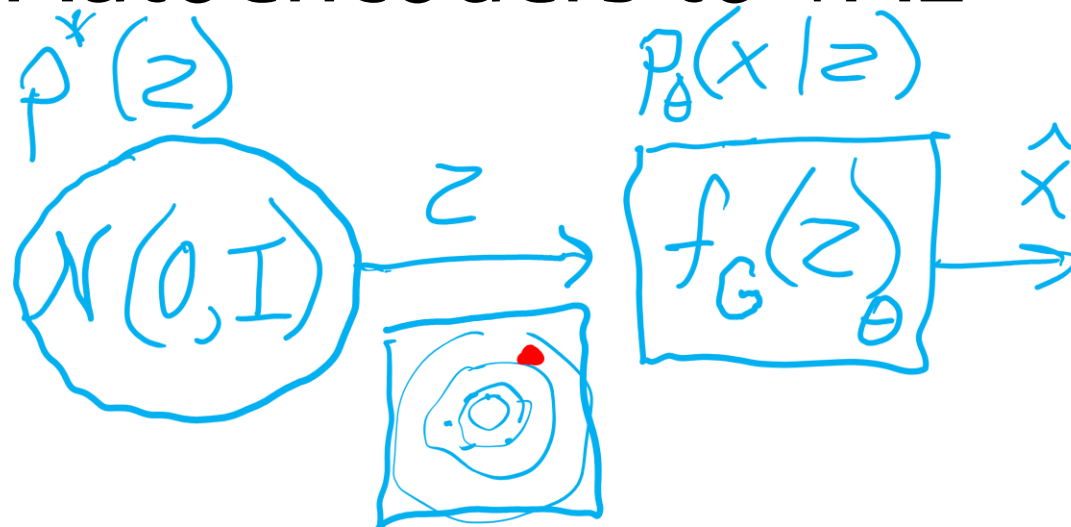
Sampling $p(x)$ from Autoencoder

How do we sample a new image?



From Autoencoders to VAE

Want



VAE Objective

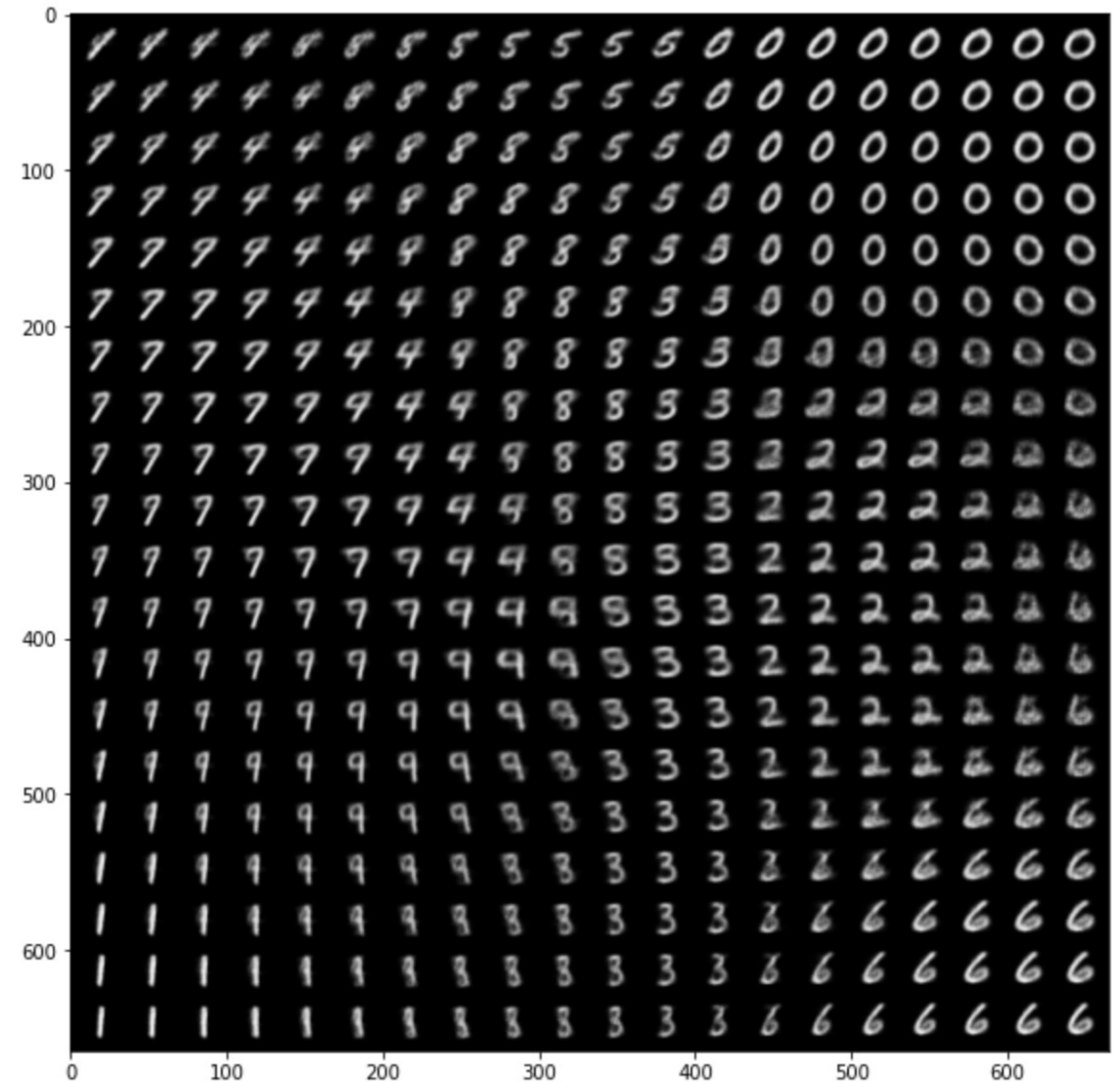
VAE Objective

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}) \right]$$

$$\tilde{\mathcal{L}}^B(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \frac{1}{L} \sum_{l=1}^L (\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}))$$

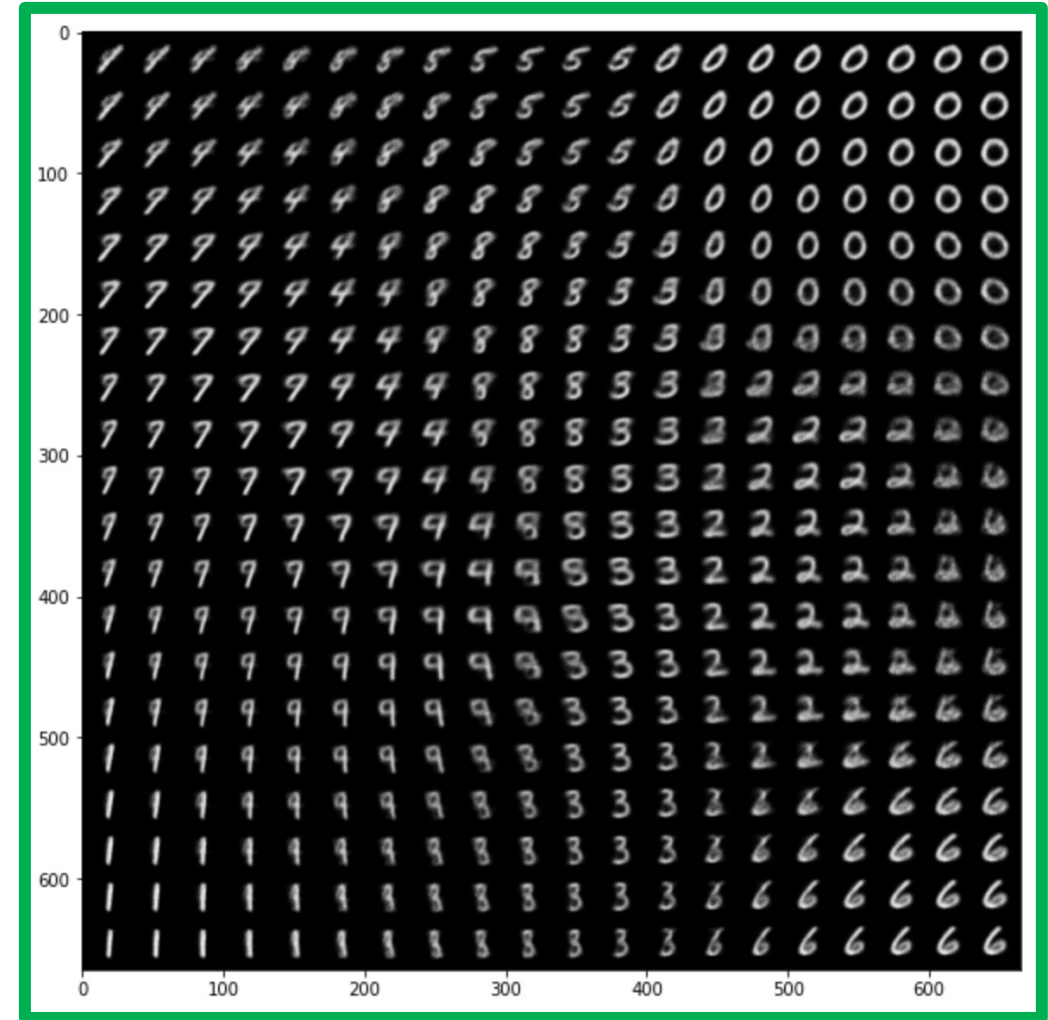
VAE Demo

- Zhuoyue Lyu, Safinah Ali, and Cynthia Breazeal. EAAI 2022.
- <https://colab.research.google.com/github/ZhuoyueLyu/5046225a9ae3675cf633c1df5f63be06/digits-interpolation-notebook-eaai.ipynb>



VAE Demo

- Feature space interpolation



Recall...

KL DIVERGENCE

KL Divergence

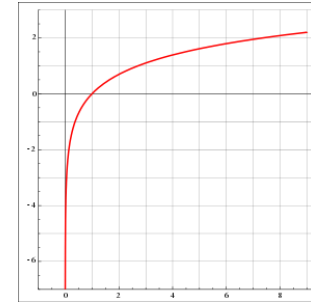
- Definition: for two distributions $q(x)$ and $p(x)$ over $x \in \mathcal{X}$, the **KL Divergence** is:

$$\text{KL}(q||p) = E_{q(x)} \left[\log \frac{q(x)}{p(x)} \right] = \begin{cases} \sum_x q(x) \log \frac{q(x)}{p(x)} \\ \int_x q(x) \log \frac{q(x)}{p(x)} dx \end{cases}$$

- Properties:
 - $\text{KL}(q || p)$ measures the **proximity** of two distributions q and p
 - KL is **not** symmetric: $\text{KL}(q || p) \neq \text{KL}(p || q)$
 - KL is minimized when $q(x) = p(x)$ for all $x \in \mathcal{X}$

$$\text{KL}(q||p) = E_{q(x)} \left[\log \frac{q(x)}{p(x)} \right]$$

KL Divergence



Understanding the Behavior of KL as an objective function

Example 1: Keeping all else constant, consider the effect of a particular x' on $\text{KL}(q || p)$

x'	$q(x')$	$p(x')$	$q(x') \log(q(x')/p(x'))$	effect on $\text{KL}(q p)$
1	0.9	0.9	0	no increase
2	0.9	0.1	1.97	big increase
3	0.1	0.9	-0.21	little decrease
4	0.1	0.1	0	little decrease

KL **does** insist on good approximations for values that have **high** probability in q

KL **does not** insist on good approximations for values that have **low** probability in q

Example 2: Which q distribution minimizes $\text{KL}(q || p)$?

$$\mathbf{p} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad
 \mathbf{q}^{(1)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad
 \mathbf{q}^{(2)} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad
 \mathbf{q}^{(3)} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$$

Q: If we're minimizing KL, why not return $q^{(3)}$?
 A: Because it's not a distribution!

$$\text{KL}(q||p) = E_{q(x)} \left[\log \frac{q(x)}{p(x)} \right]$$

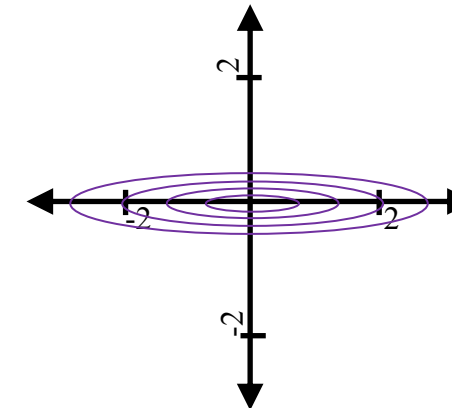
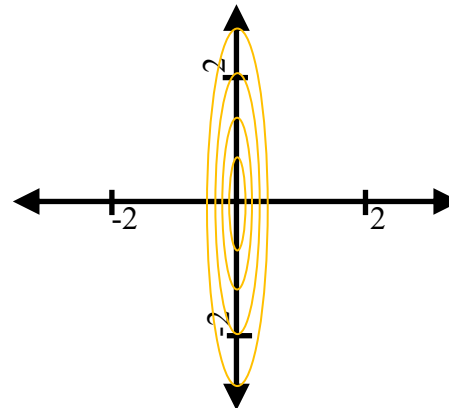
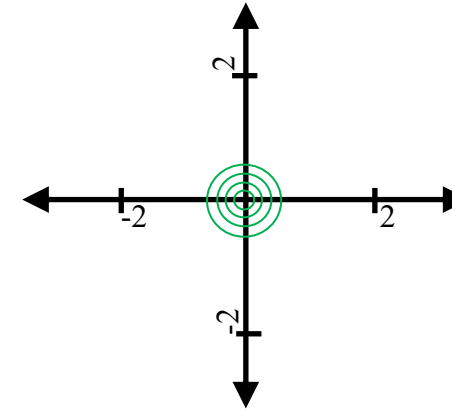
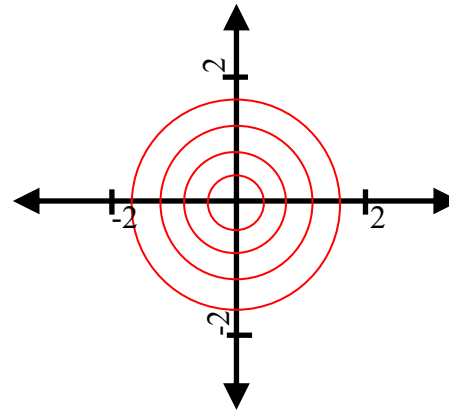
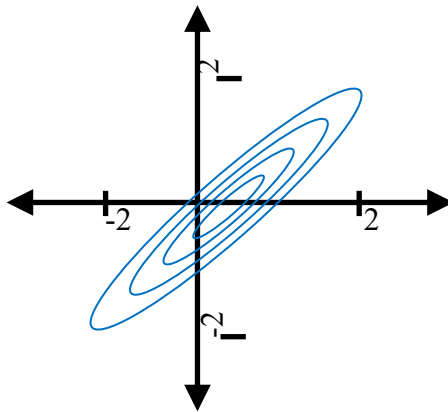
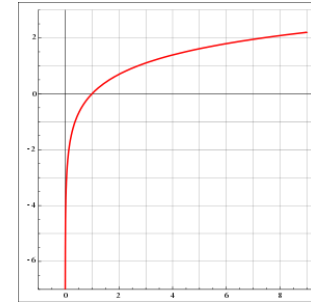
KL Divergence

Understanding the Behavior of KL as an objective function

Example 3: Which q distribution minimizes $\text{KL}(q || p)$?

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu} = [0, 0]^T, \boldsymbol{\Sigma})$$

$$q(x_1, x_2) = \mathcal{N}_1(x_1 | \mu_1, \sigma_1^2) \mathcal{N}_2(x_2 | \mu_2, \sigma_2^2)$$



Optimizing KL Divergence

- Question: How do we minimize KL?

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \underbrace{\text{KL}(q_{\theta}(\mathbf{z}) \parallel p_{\alpha}(\mathbf{z} \mid \mathbf{x}))}$$

- Answer #1: Oh no! We can't even compute this KL.

Why we can't compute KL...

$$\begin{aligned} \text{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) &= E_{q(\mathbf{z})} \left[\log \left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right) \right] \\ &= E_{q(\mathbf{z})} [\log q(\mathbf{z})] - E_{q(\mathbf{z})} [\log p(\mathbf{z} \mid \mathbf{x})] \\ &= E_{q(\mathbf{z})} [\log q(\mathbf{z})] - E_{q(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z})] + E_{q(\mathbf{z})} [\log p(\mathbf{x})] \\ &= E_{q(\mathbf{z})} [\log q(\mathbf{z})] - E_{q(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z})] + \underbrace{\log p(\mathbf{x})} \end{aligned}$$

we have the same problem
with an intractable data
likelihood $p(\mathbf{x})$ or an intractable
partition function $Z(\mathbf{x})$

we assumed this
is intractable to
compute!

Optimizing KL Divergence

- Question: How do we minimize KL?

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \underbrace{\text{KL}(q_{\theta}(\mathbf{z}) \parallel p_{\alpha}(\mathbf{z} \mid \mathbf{x}))}$$

- Answer #1: Oh no! We can't even compute this KL.

Why we can't compute KL...

$$\begin{aligned} \text{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) &= E_{q(\mathbf{z})} \left[\log \left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right) \right] \\ &= E_{q(\mathbf{z})} [\log q(\mathbf{z})] - E_{q(\mathbf{z})} [\log p(\mathbf{z} \mid \mathbf{x})] \\ &= E_{q(\mathbf{z})} [\log q(\mathbf{z})] - E_{q(\mathbf{z})} [\log \tilde{p}(\mathbf{z} \mid \mathbf{x})] + E_{q(\mathbf{z})} [\log Z(\mathbf{x})] \\ &= E_{q(\mathbf{z})} [\log q(\mathbf{z})] - E_{q(\mathbf{z})} [\log \tilde{p}(\mathbf{z} \mid \mathbf{x})] + \log Z(\mathbf{x}) \end{aligned}$$

we have the same problem
with an intractable data
likelihood $p(\mathbf{x})$ or an intractable
partition function $Z(\mathbf{x})$

we assumed this
is intractable to
compute!

Optimizing KL Divergence

- Question: How do we minimize KL?

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \underbrace{\text{KL}(q_{\theta}(\mathbf{z}) \parallel p_{\alpha}(\mathbf{z} \mid \mathbf{x}))}_{\text{KL Divergence}}$$

- Answer #2: We don't need to compute this KL

We can instead maximize the ELBO (i.e. Evidence Lower **B**ound)

$$\text{ELBO}(q_{\theta}) = E_{q_{\theta}(\mathbf{z})} [\log p_{\alpha}(\mathbf{x}, \mathbf{z})] - E_{q_{\theta}(\mathbf{z})} [\log q_{\theta}(\mathbf{z})]$$

The ELBO for a DGM

Here is why...

$$\begin{aligned} \theta &= \operatorname{argmin}_{\theta} \text{KL}(q_{\theta}(\mathbf{z}) \parallel p_{\alpha}(\mathbf{z} \mid \mathbf{x})) \\ &= \operatorname{argmin}_{\theta} E_{q_{\theta}(\mathbf{z})} [\log q_{\theta}(\mathbf{z})] - E_{q_{\theta}(\mathbf{z})} [\log p_{\alpha}(\mathbf{x}, \mathbf{z})] + \underbrace{\log p_{\alpha}(\mathbf{x})}_{\text{dropping the intractable term gives the ELBO}} \\ &= \operatorname{argmin}_{\theta} E_{q_{\theta}(\mathbf{z})} [\log q_{\theta}(\mathbf{z})] - E_{q_{\theta}(\mathbf{z})} [\log p_{\alpha}(\mathbf{x}, \mathbf{z})] \\ &= \operatorname{argmax}_{\theta} \text{ELBO}(q_{\theta}) \end{aligned}$$

Optimizing KL Divergence

- Question: How do we minimize KL?

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \underbrace{\text{KL}(q_{\theta}(\mathbf{z}) \parallel p_{\alpha}(\mathbf{z} \mid \mathbf{x}))}_{\text{KL Divergence}}$$

- Answer #2: We don't need to compute this KL

We can instead maximize the ELBO (i.e. Evidence Lower **B**ound)

$$\text{ELBO}(q_{\theta}) = E_{q_{\theta}(\mathbf{z})} [\log \tilde{p}_{\alpha}(\mathbf{z} \mid \mathbf{x})] - E_{q_{\theta}(\mathbf{z})} [\log q_{\theta}(\mathbf{z})]$$

The ELBO for a UGM

Here is why...

$$\begin{aligned} \theta &= \operatorname{argmin}_{\theta} \text{KL}(q_{\theta}(\mathbf{z}) \parallel p_{\alpha}(\mathbf{z} \mid \mathbf{x})) \\ &= \operatorname{argmin}_{\theta} E_{q_{\theta}(\mathbf{z})} [\log q_{\theta}(\mathbf{z})] - E_{q_{\theta}(\mathbf{z})} [\log \tilde{p}_{\alpha}(\mathbf{z} \mid \mathbf{x})] + \underbrace{\log Z_{\alpha}(\mathbf{x})}_{\text{dropping the intractable term gives the ELBO}} \\ &= \operatorname{argmin}_{\theta} E_{q_{\theta}(\mathbf{z})} [\log q_{\theta}(\mathbf{z})] - E_{q_{\theta}(\mathbf{z})} [\log \tilde{p}_{\alpha}(\mathbf{z} \mid \mathbf{x})] \\ &= \operatorname{argmax}_{\theta} \text{ELBO}(q_{\theta}) \end{aligned}$$

ELBO as Objective Function

What does maximizing $\text{ELBO}(q_\theta)$ accomplish?

$$\text{ELBO}(q_\theta) = E_{q_\theta(\mathbf{z})} [\log p_\alpha(\mathbf{x}, \mathbf{z})] - E_{q_\theta(\mathbf{z})} [\log q_\theta(\mathbf{z})]$$

1. The first expectation is high if q_θ puts probability mass on the same values of \mathbf{z} that p_α puts probability mass

2. The second term is the entropy of q_θ and the entropy will be high if q_θ spreads its probability mass evenly

ELBO's relation to $\log p(x)$

Theorem:

Proof #2:

Proof #1:

Key Takeaway:

ELBO's relation to $\log p(x)$

Theorem:

for any q , $\log p(x) \geq \text{ELBO}(q)$
i.e. $\text{ELBO}(q)$ is a lower bound on $\log p(x)$

Proof #2:

- ① $\log p(x) = \text{KL}(q||p) + \text{ELBO}(q)$
- ② $\text{KL}(q||p) \geq 0$ (without proof)
- ③ $\Rightarrow \log p(x) \geq \text{ELBO}(q)$

Proof #1:

Recall Jensen's Inequality: $f(E[x]) \geq E[f(x)]$, for concave f

$$\begin{aligned}\log p(x) &= \log \int_{\mathbf{z}} p(x, \mathbf{z}) d\mathbf{z} \quad (\text{marginal}) \\ &= \log \int_{\mathbf{z}} p(x, \mathbf{z}) \frac{q(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \quad (\text{mult. by 1}) \\ &= \log E_{q(\mathbf{z})} \left[\frac{p(x, \mathbf{z})}{q(\mathbf{z})} \right] \quad (\text{def. of expectation}) \\ &\geq E_{q(\mathbf{z})} \left[\log \left(\frac{p(x, \mathbf{z})}{q(\mathbf{z})} \right) \right] \quad (\text{by Jensen's Ineq.}) \\ &= E_{q(\mathbf{z})} [\log p(x, \mathbf{z})] - E_{q(\mathbf{z})} [\log q(\mathbf{z})] = \text{ELBO}(q) \\ \Rightarrow \log p(x) &\geq \text{ELBO}(q)\end{aligned}$$

Key Takeaway:

minimizing KL is the same as
finding a tight $\text{ELBO}(q)$ lower bound for $\log p(x)$

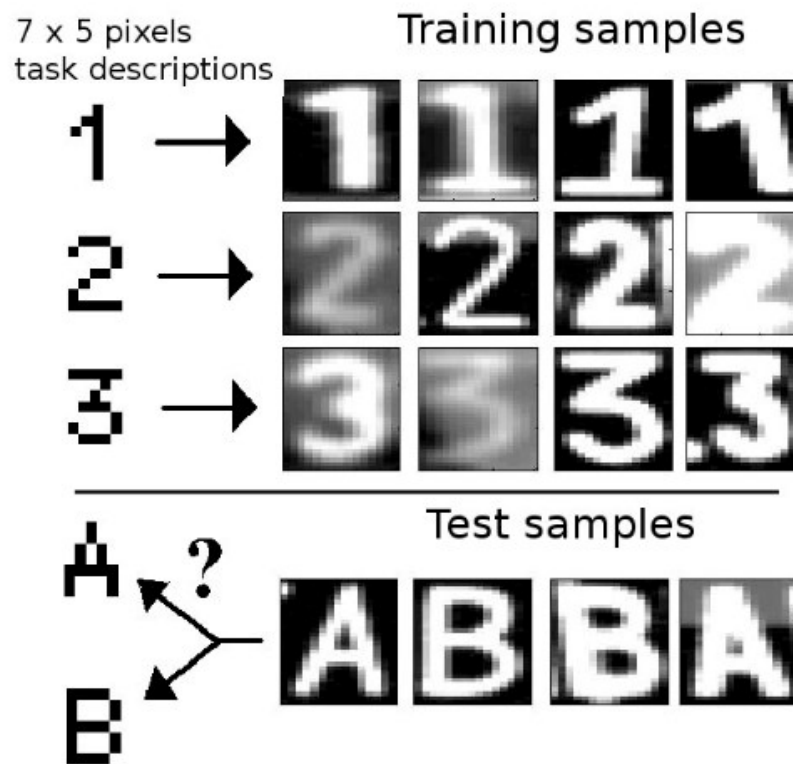
ZERO-SHOT AND FEW-SHOT LEARNING

Zero-shot Learning

- Definition:** in **zero-shot learning** we assume that training data does not contain any examples of the labels that appear in the test data

$z \rightarrow$	1	2	3	4	5
$d(z) \rightarrow$	1	2	3	A	B
x_t	y_t^1	y_t^2	y_t^3	y_t^4	y_t^5
1	1	0	0	-	-
2	0	1	0	-	-
3	0	0	1	-	-
A	-	-	-	1	0
B	-	-	-	0	1

⏟
⏟
 training data test data



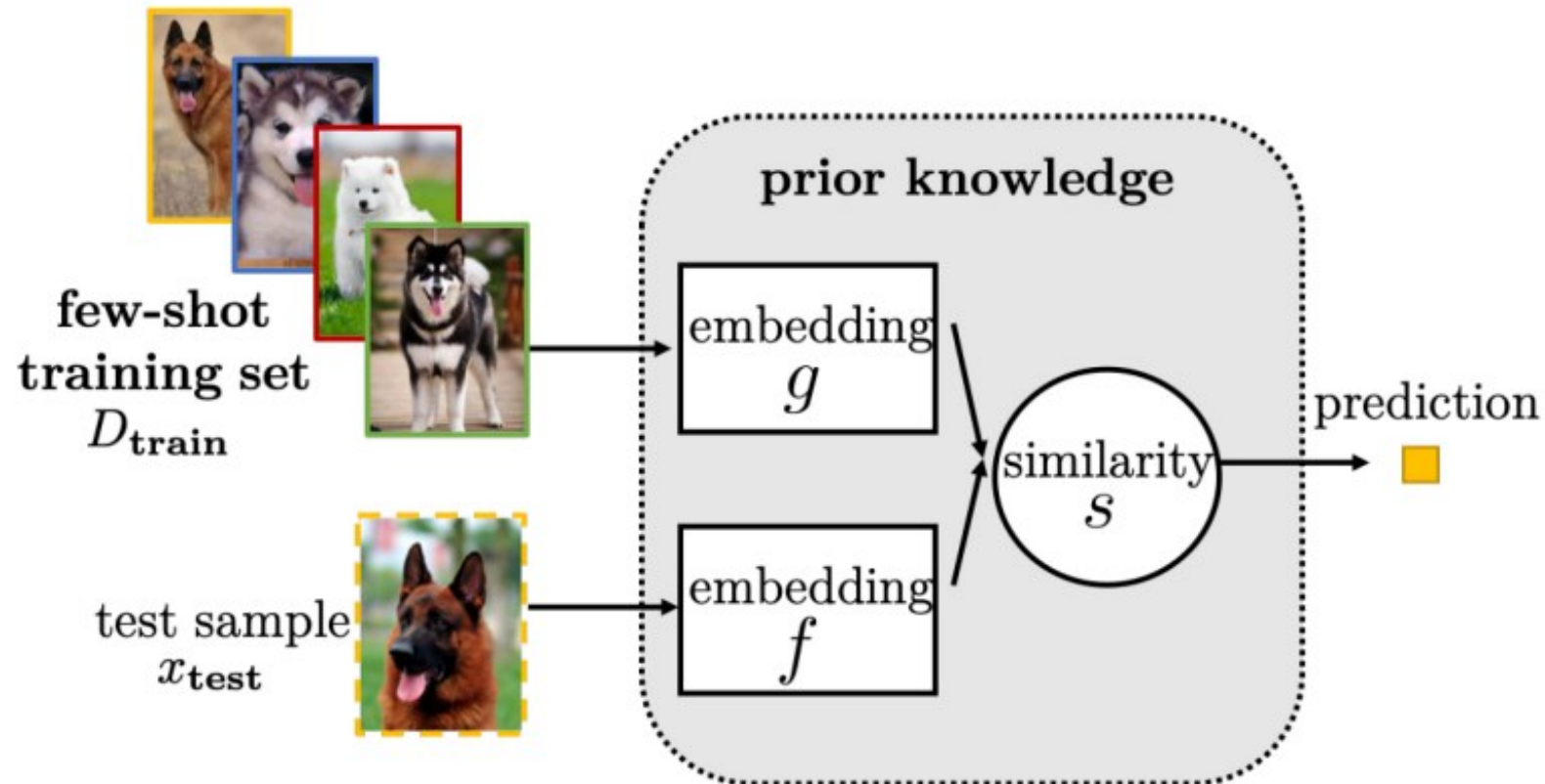
Question:

How can we hope to learn in a setting where we have no labeled examples?

Answer:

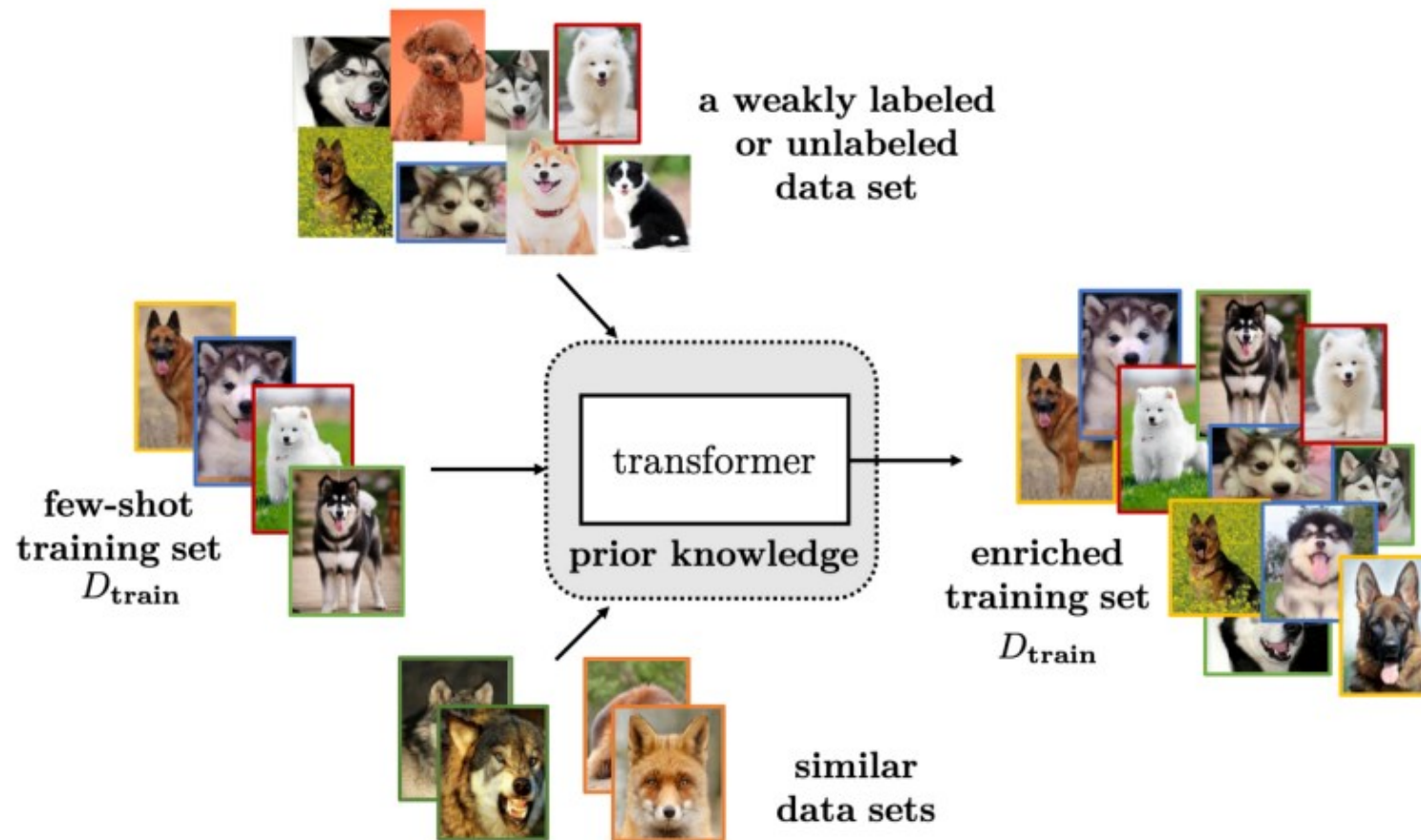
Few-shot Learning

- **Definition:** in **few-shot learning** we assume that training data contains a handful (maybe two, three, or four) examples of each label



Few-shot Learning

- **Definition:** in **few-shot learning** we assume that training data contains a handful (maybe two, three, or four) examples of each label



PROMPTING

Prompting

- Language models are trained to maximize the likelihood of sequences in the training data
- Most condition on the previous tokens to generate the next tokens

Definition: An *autoregressive language model* defines a probability distribution over sequences $\mathbf{x}_{1:T}$ of the form:

$$p(\mathbf{x}_{1:T}) = \prod_{t=1}^T p(x_t \mid x_1, \dots, x_{t-1})$$

- Key idea behind prompting: provide a prefix string to the such that its likely completion is the answer you want

Prompting Examples

GPT-3

Prompt
(i.e. context)

The City
BY C. P. CAVAFY
TRANSLATED BY EDMUND KEELEY
[Poem text omitted]

SOME TREES
John Ashbery
[Poem text omitted]

Shadows on the Way
Wallace Stevens

Output
(i.e. generated
poem)

I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That color is my friend
And purple must surround me too

The yellow of the sun is no more
Intrusive than the bluish snow
That falls on all of us. I must have
Grey thoughts and blue thoughts walk with me
If I am to go away at all.

Conditional Sampling:

To draw a sample, feed in a prompt (i.e. context) and then sample from the model conditioned on that prompt

translation

Prompting Examples

Llama-2-70B

Prompt

El último dueño de esta propiedad había sido un hombre soltero, que alcanzó una muy avanzada edad, y que durante gran parte de su existencia tuvo en su hermana una fiel compañera y ama de casa.

English translation:

Model Output

The last owner of this property was a bachelor man who lived to a very advanced age, and who had his sister as a faithful companion and housekeeper for most of his life.

Conditional Sampling:

To draw a sample, feed in a prompt (i.e. context) and then sample from the model conditioned on that prompt

English source from Sense & Sensibility

The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister.

Prompting Examples

Llama-2-70B

Prompt

Rhodochiton is a genus of flowering plants within the family Plantaginaceae, native to southern Mexico and neighbouring Guatemala. They climb by means of twining leaf stalks. One of the three species, "Rhodochiton atrosanguineus", the purple bell vine, is grown as an ornamental plant. All three species are sometimes included in "Lophospermum".

Question: You can find the purple bell vine in more than one country. True, False, or Neither?

Model Output

Answer: True.

Explanation: The purple bell vine (Rhodochiton atrosanguineus) is native to southern Mexico and neighbouring Guatemala, which means it can be found in more than one country.

Correct Answer

True.

Prompting Examples

Llama-2-70B

Prompt

Story: Jason listened to the weather and heard it was going to be sunny. He thought the kids might like to go swimming. He gathered up the swimsuits, towels and sunscreen. Jason and the kids got into the truck and drove to the beach. They spent the next 2 hours playing and splashing in the surf.

One-sentence Summary:

Model Output

Jason took the kids swimming at the beach after hearing the weather forecast, gathering necessary items and driving there.

What else can a large LM (attempt to) do?

Using the idea of **prompts**, we can apply LMs to a **variety of different problems** in natural language processing.

In the **zero-shot setting**, we simply feed the context to the model and observe how it completes the sequence. (i.e. there is no additional training)

Answer fact-based questions:

Context → Organisms require energy in order to do what?

Correct Answer → mature and develop.
Incorrect Answer → rest soundly.
Incorrect Answer → absorb light.
Incorrect Answer → take in nutrients.

Complete sentences logically:

Context → My body cast a shadow over the grass because

Correct Answer → the sun was rising.
Incorrect Answer → the grass was cut.

Complete analogies:

Context → lull is to trust as

Correct Answer → cajole is to compliance
Incorrect Answer → balk is to fortitude
Incorrect Answer → betray is to loyalty
Incorrect Answer → hinder is to destination
Incorrect Answer → soothe is to passion

Reading comprehension:

Context → anli 1: anli 1: Fulton James MacGregor MSP is a Scottish politician who is a Scottish National Party (SNP) Member of Scottish Parliament for the constituency of Coatbridge and Chryston. MacGregor is currently Parliamentary Liaison Officer to Shona Robison, Cabinet Secretary for Health & Sport. He also serves on the Justice and Education & Skills committees in the Scottish Parliament.
Question: Fulton James MacGregor is a Scottish politician who is a Liaison officer to Shona Robison who he swears is his best friend. True, False, or Neither?

Correct Answer → Neither
Incorrect Answer → True
Incorrect Answer → False

Zero-shot LLMs

- GPT-2 (1.5B parameters) for unsupervised prediction on various tasks
- GPT-2 models $p(\text{output} \mid \text{input}, \text{task})$
 - translation: (*translate to french, english text, french text*)
 - reading comprehension: (*answer the question, document, question, answer*)
- Why does this work?

“I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**”

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: “**Mentez mentez, il en restera toujours quelque chose,**” which translates as, “**Lie lie and something will always remain.**”

“I hate the word ‘**perfume,**’” Burr says. ‘It’s somewhat better in French: ‘**parfum.**’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “-**Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “- **How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Zero-shot LLMs

- GPT-2 (1.5B parameters) for unsupervised prediction on various tasks
- GPT-2 models $p(\text{output} \mid \text{input, task})$
 - translation: (*translate to french, english text, french text*)
 - reading comprehension: (*answer the question, document, question, answer*)
- Why does this work?

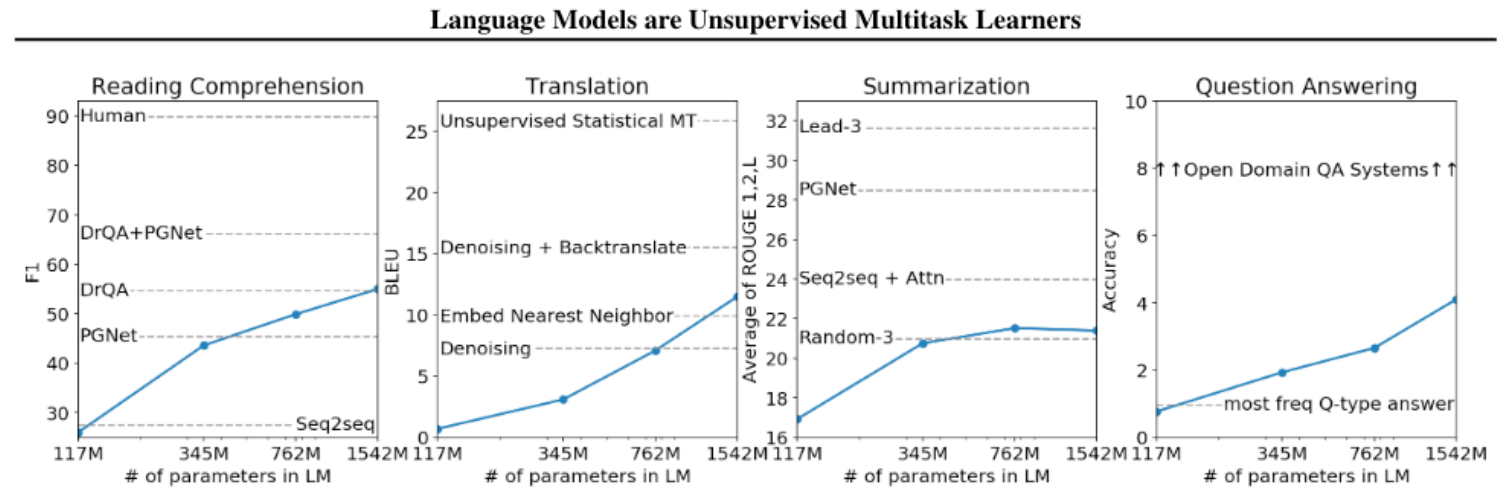


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Prompting for Instruction Fine-tuned Models

- Models like ChatGPT, Llama-2 Chat, etc. have been fine-tuned as chat assistants
- These (often) were trained with specific prompt templates that segment the prompt into different parts: (1) system (2) assistant (3) user

Llama-2 Chat

```
sys: [INST] <<SYS>>  
You are a helpful AI assistant...  
<</SYS>> [/INST]  
  
user: [INST]  
Organisms require energy in order to do what?  
[/INST]  
  
asst: mature and develop
```

Alpaca

```
sys: ### Instruction:  
You are a helpful AI assistant...  
  
user: ### Instruction:  
Organisms require energy in order to do what?  
  
asst: ### Response:  
mature and develop
```

Prompting for Instruction Fine-tuned Models

Llama-2-70B

Prompt

Story: Jason listened to the weather and heard it was going to be sunny. He thought the kids might like to go swimming. He gathered up the swimsuits, towels and sunscreen. Jason and the kids got into the truck and drove to the beach. They spent the next 2 hours playing and splashing in the surf.

One-sentence Summary:

Model Output

Jason took the kids swimming at the beach after hearing the weather forecast, gathering necessary items and driving there.

Llama-2-7B Chat

Story: Jason listened to the weather and heard it was going to be sunny. He thought the kids might like to go swimming. He gathered up the swimsuits, towels and sunscreen. Jason and the kids got into the truck and drove to the beach. They spent the next 2 hours playing and splashing in the surf.

One-sentence Summary:

Sure! Here is a one-sentence summary of the story:
Jason and the kids went on a fun-filled day at the beach, playing and splashing in the surf after gathering all the necessary items, including swimsuits, towels, and sunscreen.

Prompting for Instruction Fine-tuned Models

Llama-2-70B

Llama-2-7B Chat

Prompt

Story: Jason listened to the weather and heard it was going to be sunny. He thought the kids might like to go swimming. He gathered up the swimsuits, towels and sunscreen. Jason and the kids got into the truck and drove to the beach. They spent the next 2 hours playing and splashing in the surf.

Story: Jason listened to the weather and heard it was going to be sunny. He thought the kids might like to go swimming. He gathered up the swimsuits, towels and sunscreen. Jason and the kids got into the truck and drove to the beach. They spent the next 2 hours playing and splashing in the surf.

Llama-2-7B Chat was *instruction fine-tuned* (more on this later) and so its responses look quite different from the Llama-2 models that were not

One-sentence Summary:

One-sentence Summary:

Model Output

Jason took the kids swimming at the beach after hearing the weather forecast, gathering necessary items and driving there.

Sure! Here is a one-sentence summary of the story:
Jason and the kids went on a fun-filled day at the beach, playing and splashing in the surf after gathering all the necessary items, including swimsuits, towels, and sunscreen.

IN-CONTEXT LEARNING

Few-shot Learning with LLMs

Suppose you have...

- a dataset $D = \{(x_i, y_i)\}_{i=1}^N$ and N is rather small (i.e. few-shot setting)
- a very large (billions of parameters) pre-trained language model

There are two ways to “learn”

This section!



Option A: Supervised fine-tuning

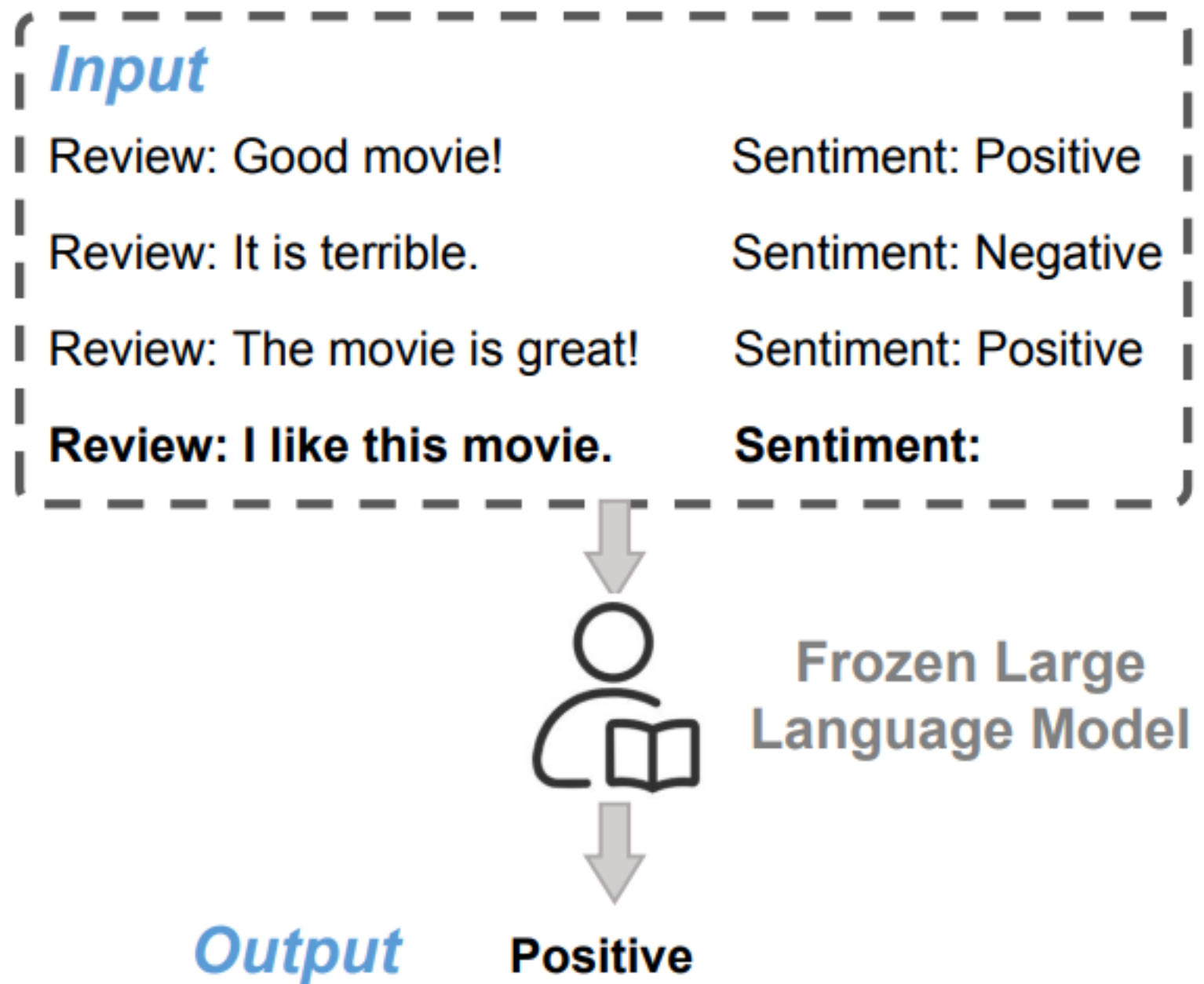
- **Definition:** fine-tune the LLM on the training data using...
 - a standard supervised objective
 - backpropagation to compute gradients
 - your favorite optimizer (e.g. Adam)
- **Pro:** fits into the standard ML recipe
- **Pro:** still works if N is large
- **Con:** backpropagation requires $\sim 3x$ the memory and computation time as the forward computation
- **Con:** you might not have access to the model weights at all (e.g. because the model is proprietary)

Option B: In-context learning

- **Definition:**
 1. feed training examples to the LLM as a prompt
 2. allow the LLM to infer patterns in the training examples during inference (i.e. decoding)
 3. take the output of the LLM following the prompt as its prediction
- **Con:** the prompt may be very long and Transformer LMs require $O(N^2)$ time/space where N = length of context
- **Pro:** no backpropagation required and only one pass through the training data
- **Pro:** does not require model weights, only API access

Few-shot In-context Learning

- Few-shot learning can be done via in-context learning
- Typically, a task description is presented first
- Then a sequence of input/output pairs from a training dataset are presented in sequence



Few-shot In-context Learning

- Few-shot learning can be done via in-context learning
- Typically, a task description is presented first
- Then a sequence of input/output pairs from a training dataset are presented in sequence

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

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

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

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 ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Few-shot In-context Learning

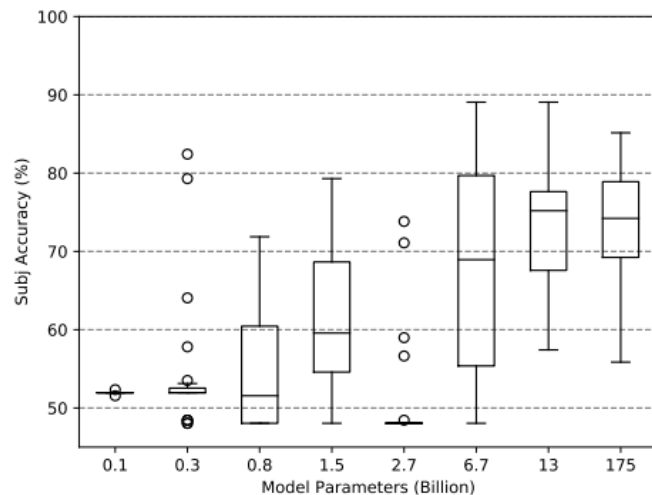
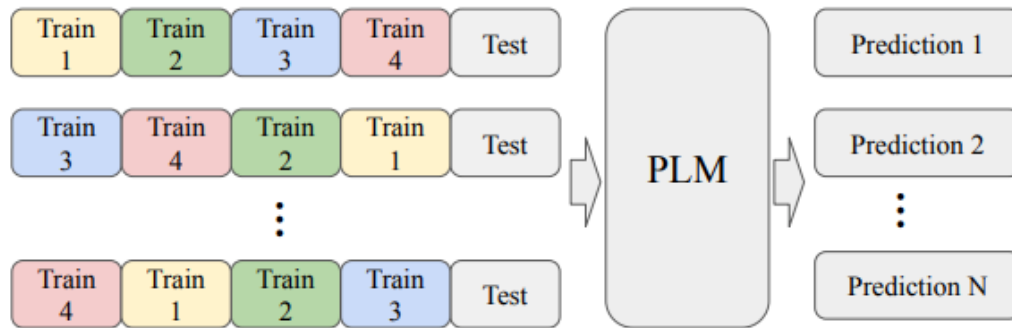


Figure 1: Four-shot performance for 24 different sample orders across different sizes of GPT-family models (GPT-2 and GPT-3) for the SST-2 and Subj datasets.

In-context learning can be sensitive to...

1. **the order the training examples are presented**
2. the balance of labels (e.g. positive vs. negative)
3. the number of unique labels covered

Few-shot In-context Learning

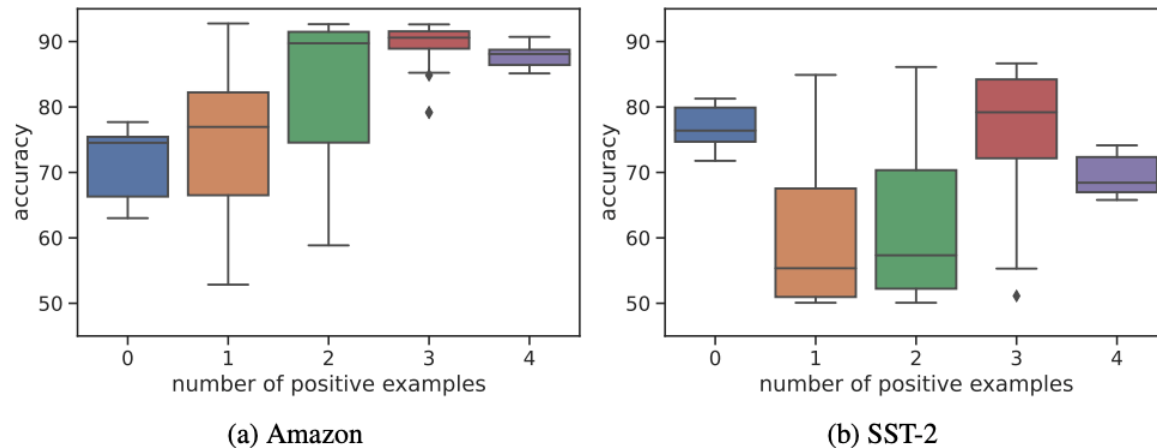
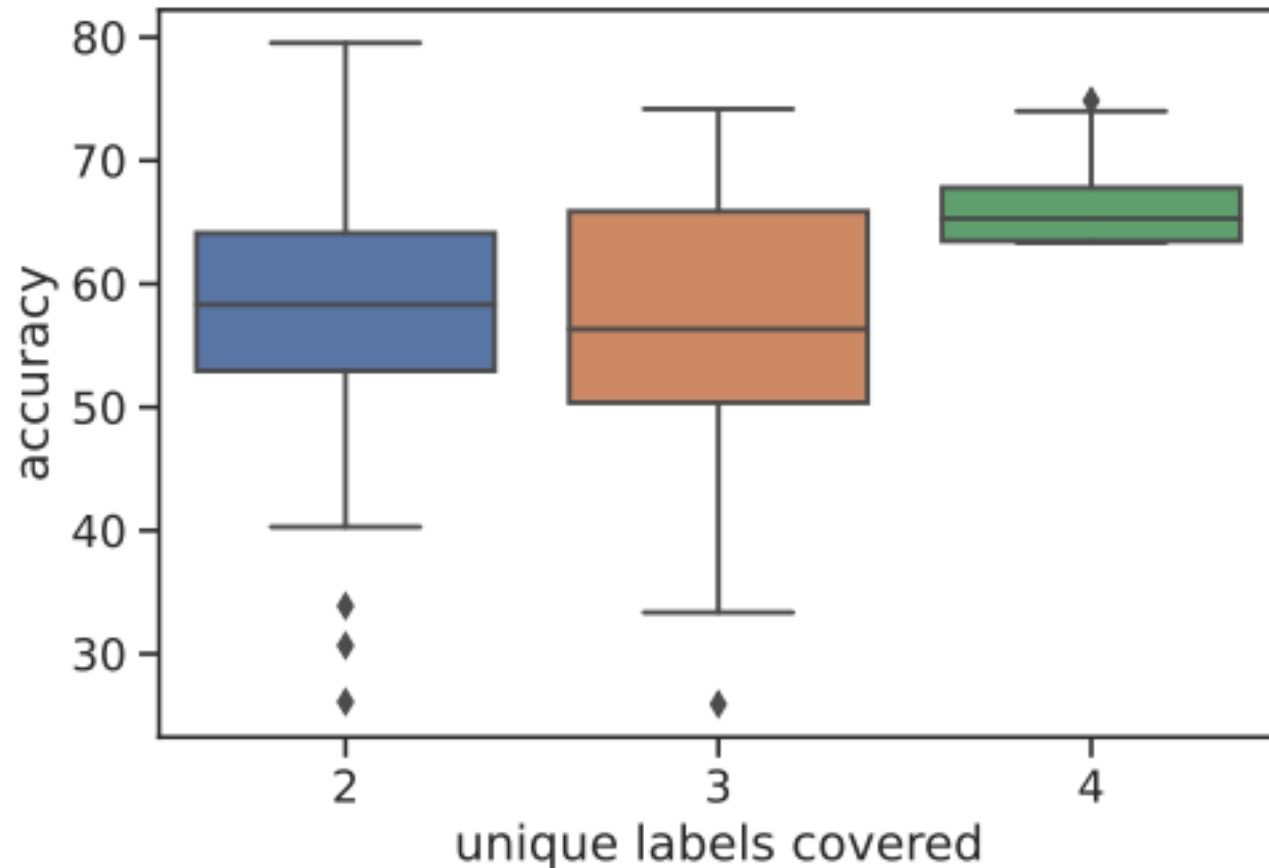


Figure 3: Accuracies of Amazon and SST-2 with varying **label balance** (number of positive examples in demonstration), across 100 total random samples of 4 demonstration examples.

In-context learning can be sensitive to...

1. the order the training examples are presented
2. **the balance of labels (e.g. positive vs. negative)**
3. the number of unique labels covered

Few-shot In-context Learning



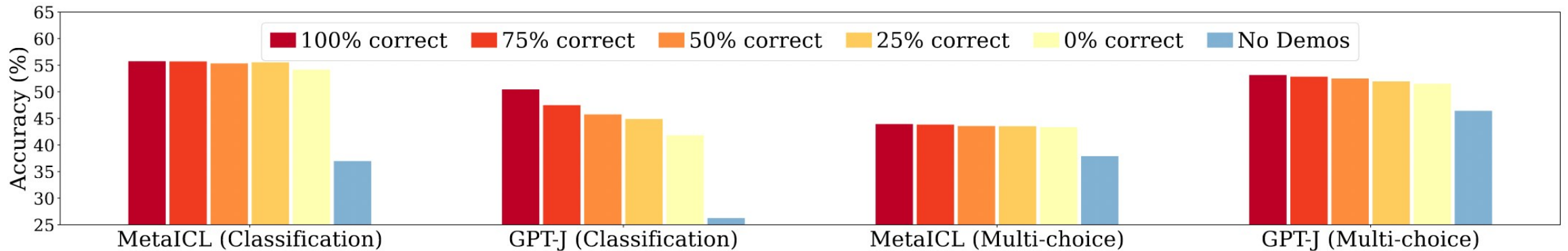
In-context learning can be sensitive to...

1. the order the training examples are presented
2. the balance of labels (e.g. positive vs. negative)
3. **the number of unique labels covered**

Few-shot In-context Learning

You would expect these to be important...

- A. whether or not the training examples have the true label (as opposed to a random one)
 - B. having more in-context training examples
- ...but it's not always the case



Few-shot In-context Learning

You would expect these to be important...

A. whether or not the training examples have the true label
(as opposed to a random one)

B. having more in-context training examples

...but it's not always the case

