Washington University in St. Louis
James McKelvey School of Engineering

CSE 561A: Large Language Models

Spring 2024

Lecture 3: Scaling up Language Models and In-Context Learning

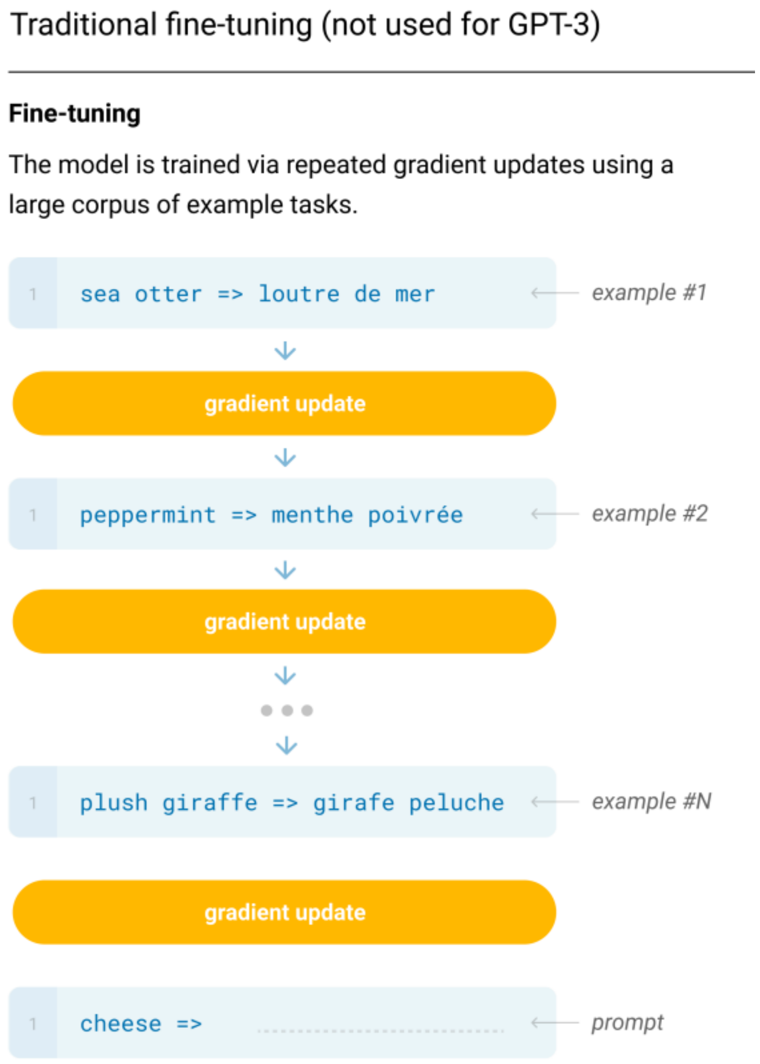
Jiaxin Huang

Content

- **Scaling up Language Model: GPT-3**
- Open-Source Model: Llama 2
- What Makes In-Context Learning Work?: Empirical Analysis
- What Makes In-Context Learning Work?: Theoretical Analysis

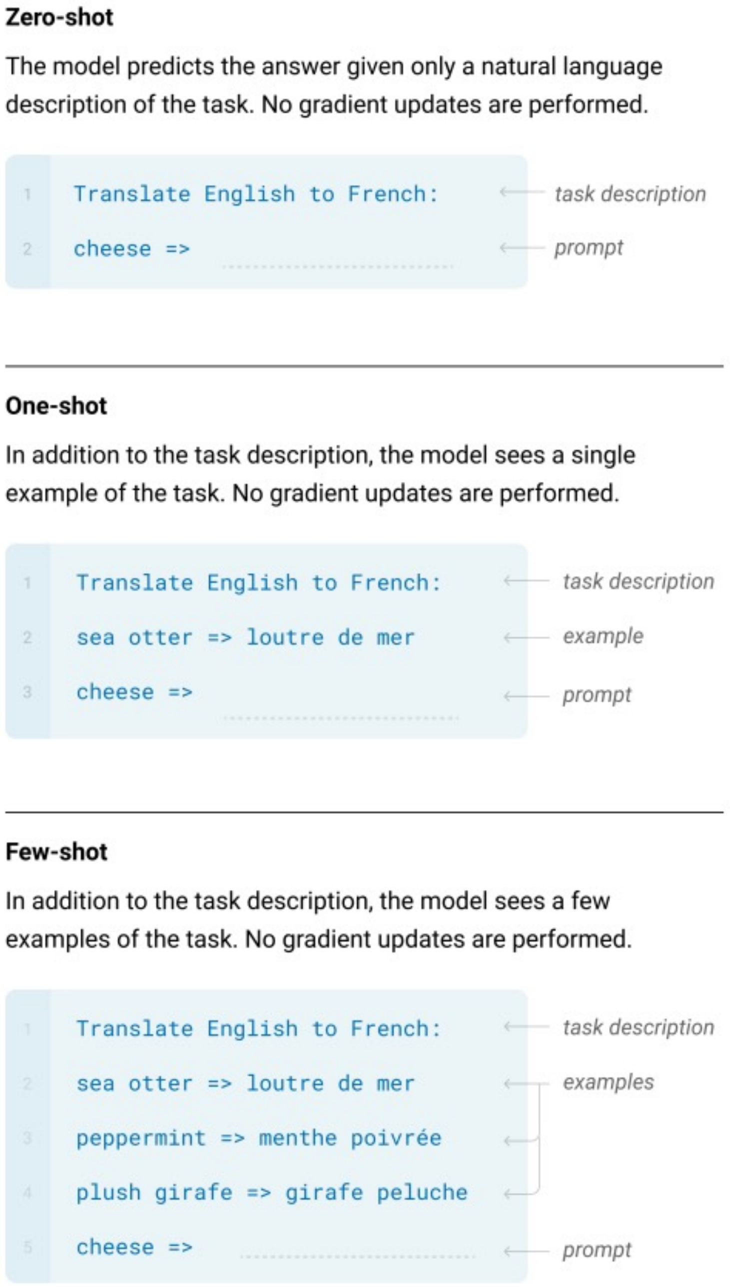
Limitations of the Fine-tuning Paradigm

- Requires a large number of labeled training examples for the down-stream task
- Hard to generalize to new tasks
- Computationally expensive when language models scale up



In-Context Learning

- Does not need model training
- Use instruction to describe the goal of a task
- Provide K-shot examples to the model at test time



In-context learning with Different Labels

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. // ________

LM

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. // ________

LM

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

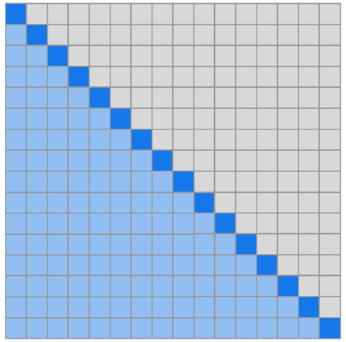
https://arxiv.org/pdf/2005.14165

Scaling up GPT Models – Architecture

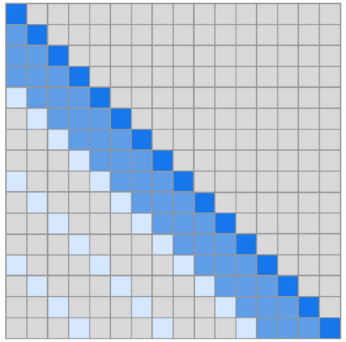
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

GPT-3 Architecture Improvement

- Sparse attention for longer context window: 1024 → 2048



Dense Attention: Tokens attend to every previous tokens



Sparse Attention: Tokens attend to sliding window

- This allows the local context and global information to propagate more efficiently

Scaling up GPT Models – Pre-Training Data

- GPT-3 is trained on ~300B tokens, compared to GPT-2 with ~40B tokens.

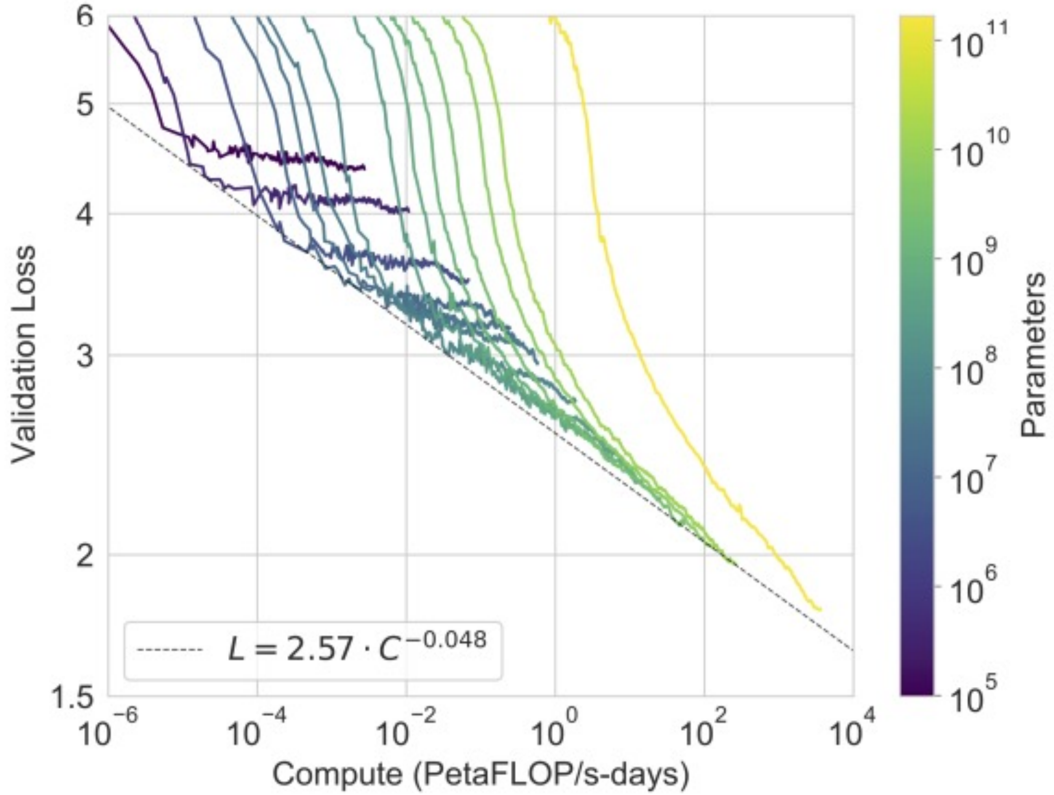
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Training objective remains the same:

$$\mathcal{L}_{\mathrm{LM}} = -\sum_i \log p(x_i \mid x_{i-k}, \ldots, x_{i-1})$$

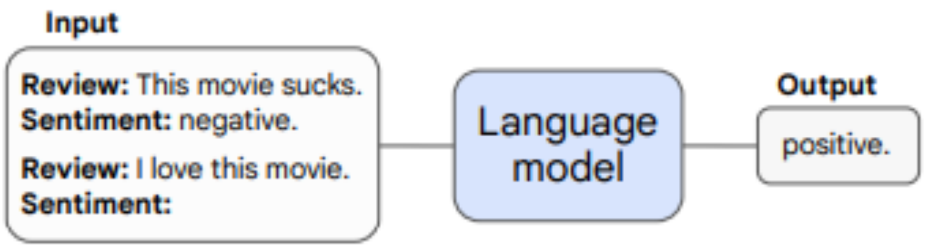
Validation Set Performance

- Performance on validation set (cross entropy loss on standard language modeling task) follows a power-law trend with respect to the amount of computation in training



Test Set Performance on Few-Shot Inference

- Example of 1-shot inference:



- As language models scale up, their one-shot/few-shot performance gradually exceeds fine-tuned smaller-sized models.

Evaluation on Question Answering Tasks

- Open-domain setting: offers external sources including the final answer
- GPT-3 answers questions without looking at the sources

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Evaluation on Reasoning Tasks

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

- GPT-3 achieves lower score than fine-tuned models.
- Reasoning process is commonly not explicitly stated in texts, so GPT-3 benefits less from the pre-training stage.

Limitations of GPT-3

- Computationally expensive
- Lack of reasoning ability
- Closed-source model

Content

- Scaling up Language Model: GPT-3
- **Open-Source Model: Llama 2**
- What Makes In-Context Learning Work?: Empiricial Analysis
- What Makes In-Context Learning Work?: Theoretical Analysis

An Open-Source Model: Llama 2

Llama 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron* Louis Martin† Kevin Stone†
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra
Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang
Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang
Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic
Sergey Edunov Thomas Scialom*

GenAI, Meta

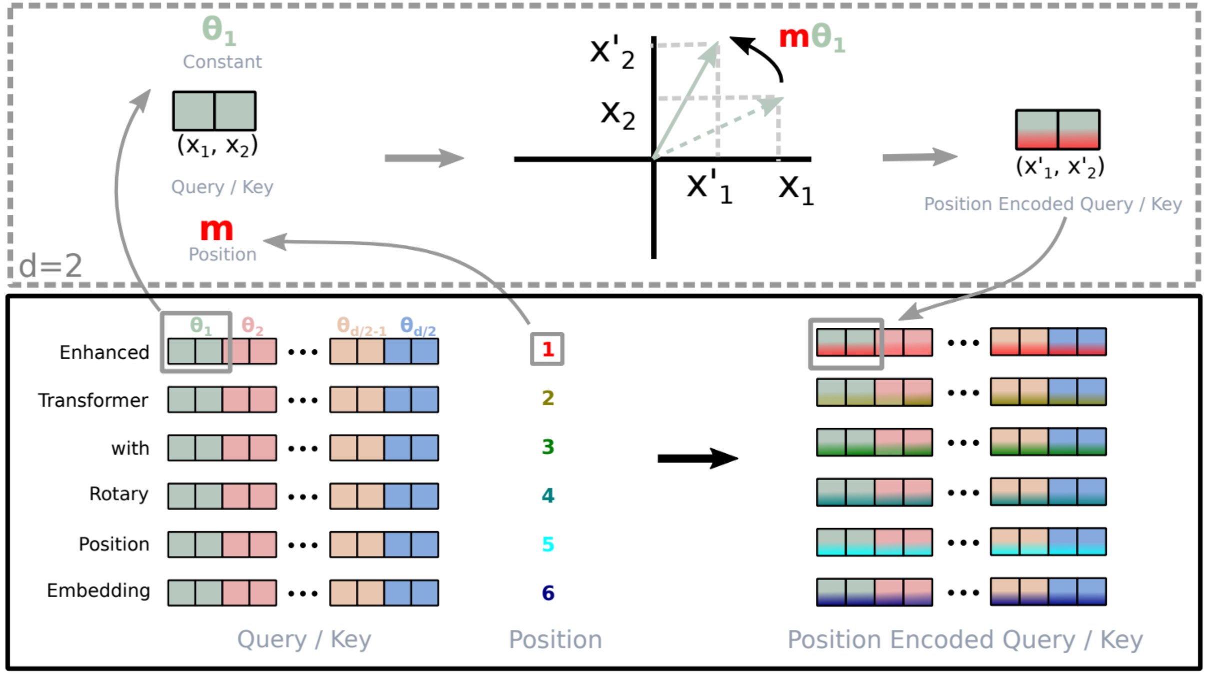
https://arxiv.org/pdf/2307.09288

Main Contribution

- Llama 2 is the first open-sourced model that matches closed sourced models' performance.
- Llama 2 is available in multiple sizes: 7B, 13B, and 70B.

Llama 2 Improvement: Rotary Position Embedding

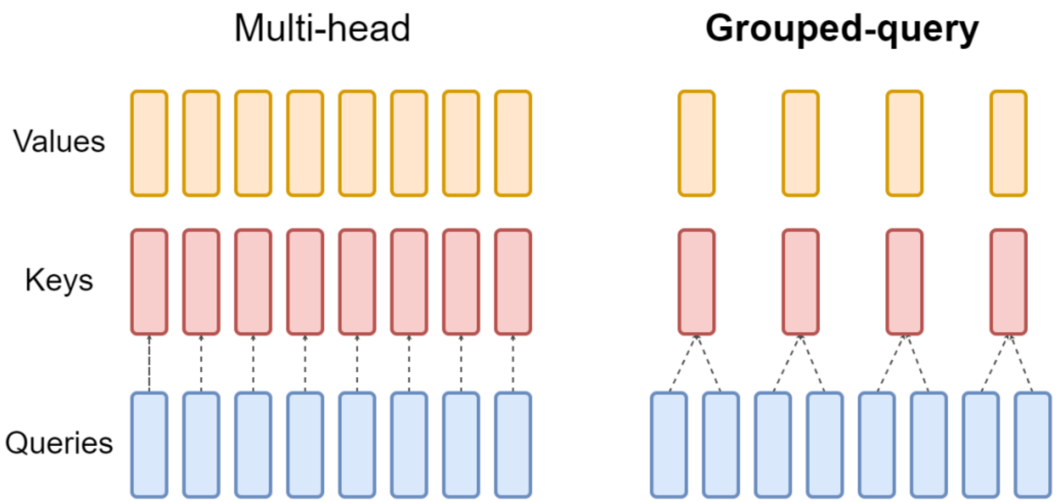
- Absolute positional encoding is simple, but may not generalize well in longer sequences
- Integrate relative position between tokens in the self-attention matrix



RoFormer: Enhanced Transformer with Rotary Position Embedding. Su et al, 2021.
https://arxiv.org/abs/2104.09864

Llama 2 Improvement: Grouped-Query Attention

- Multi-query attention has different key and value heads across all query heads.
- Grouped-query attention instead shares single key and value heads for each group of query heads.



Llama 2 Performance

- Llama 2 model is not as good as proprietary models, but still very competitive (as a pre-trained only model)

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	Llama 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	**86.1**	85.0
Natural Questions (1-shot)	–	–	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	**65.7**	51.2

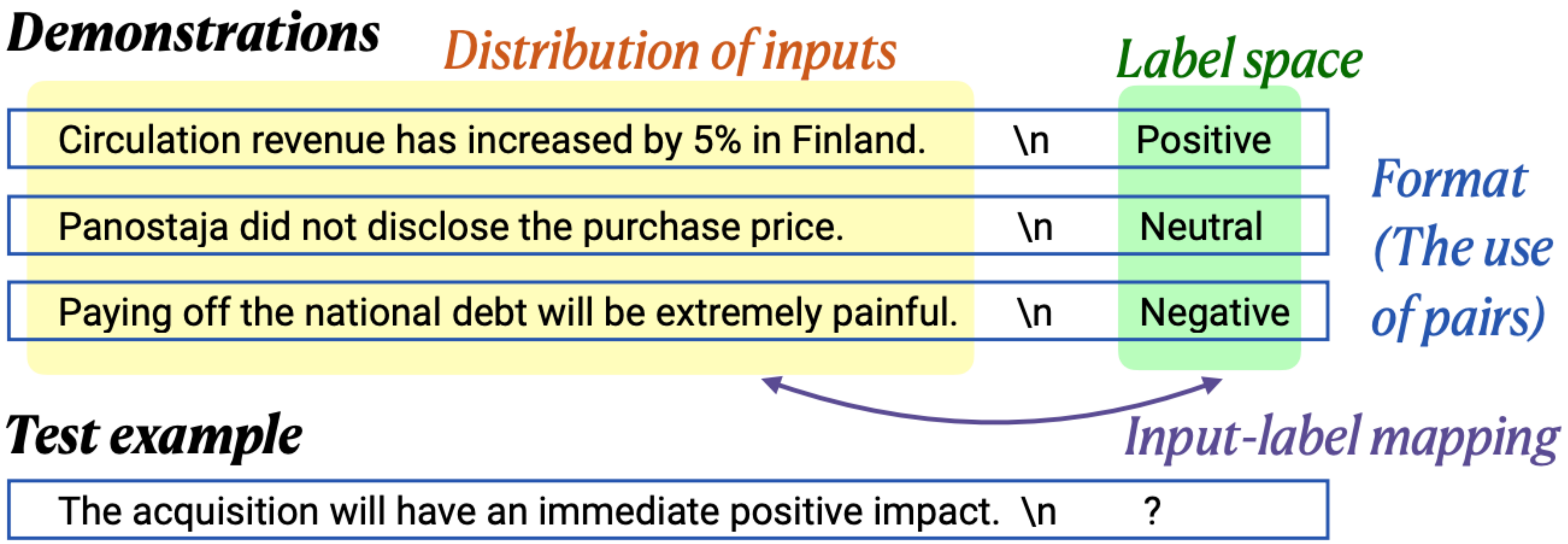
Content

- Scaling up Language Model: GPT-3
- Open-Source Model: Llama 2
- **What Makes In-Context Learning Work?: Empirical Analysis**
- What Makes In-Context Learning Work?: Theoretical Analysis

What makes in-context learning work?

- Which part of in-context learning makes it work?
- Experiment 1: replace gold labels with random labels

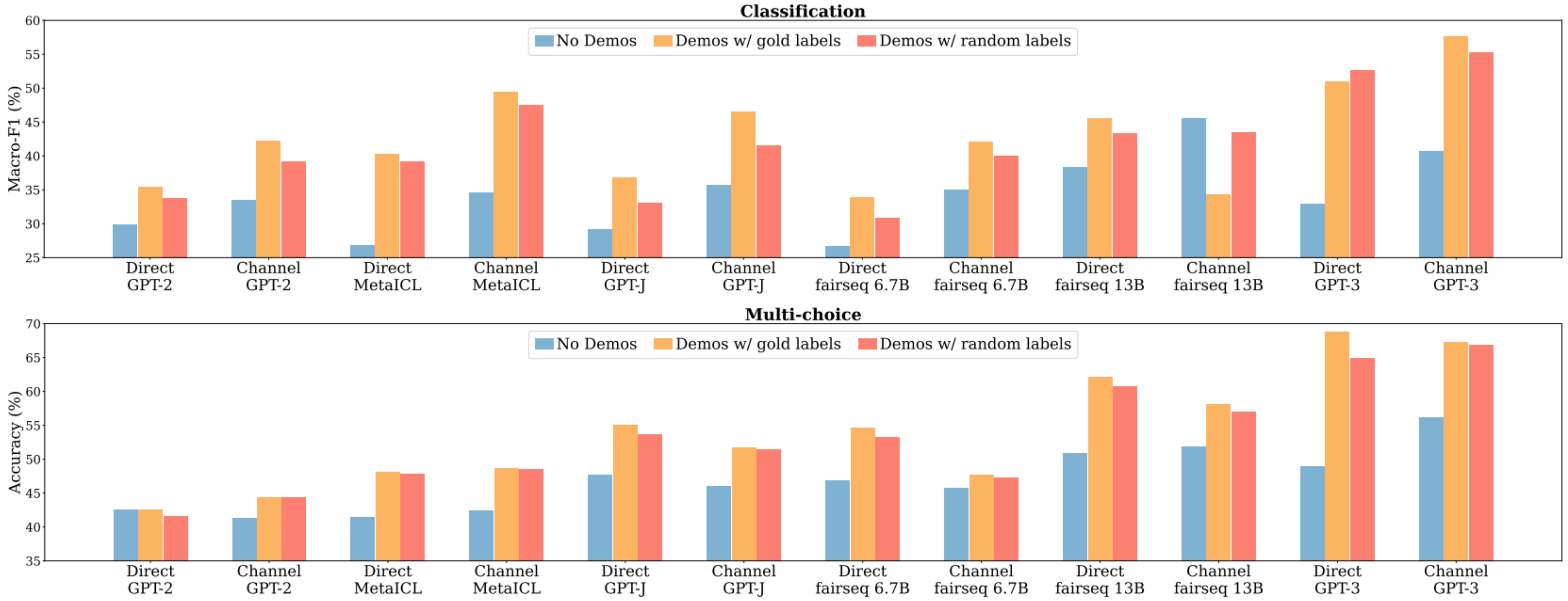
(x, y) -> (x, y')



Rethinking the Role of Demonstrations: what makes in-context learning work? Min et al. 2022.
https://arxiv.org/abs/2202.12837

Experiment 1: Replace Gold Labels with Random Labels

- Random labels only slightly hurt the performance (less than 5%)
- The model can recover the expected input labels



Experiment 2: Change Portion of Correct Labels

- Using wrong label demos is much better than no demos at all
- Using correct label demos improve the performance

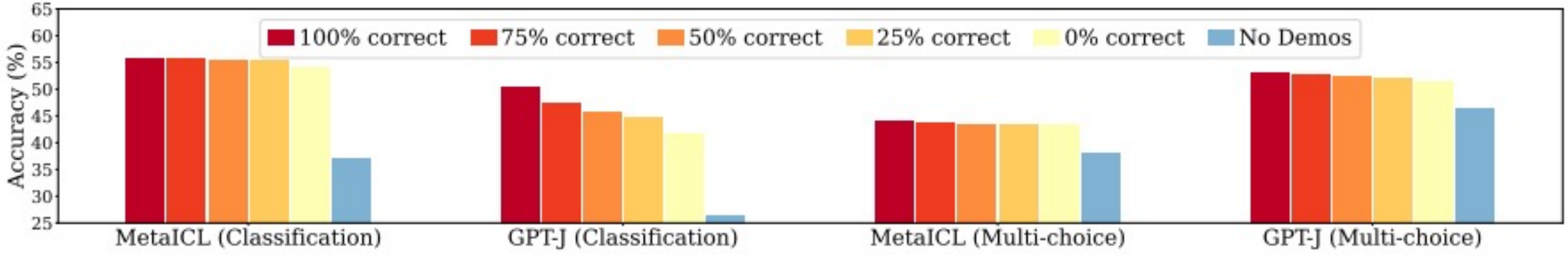
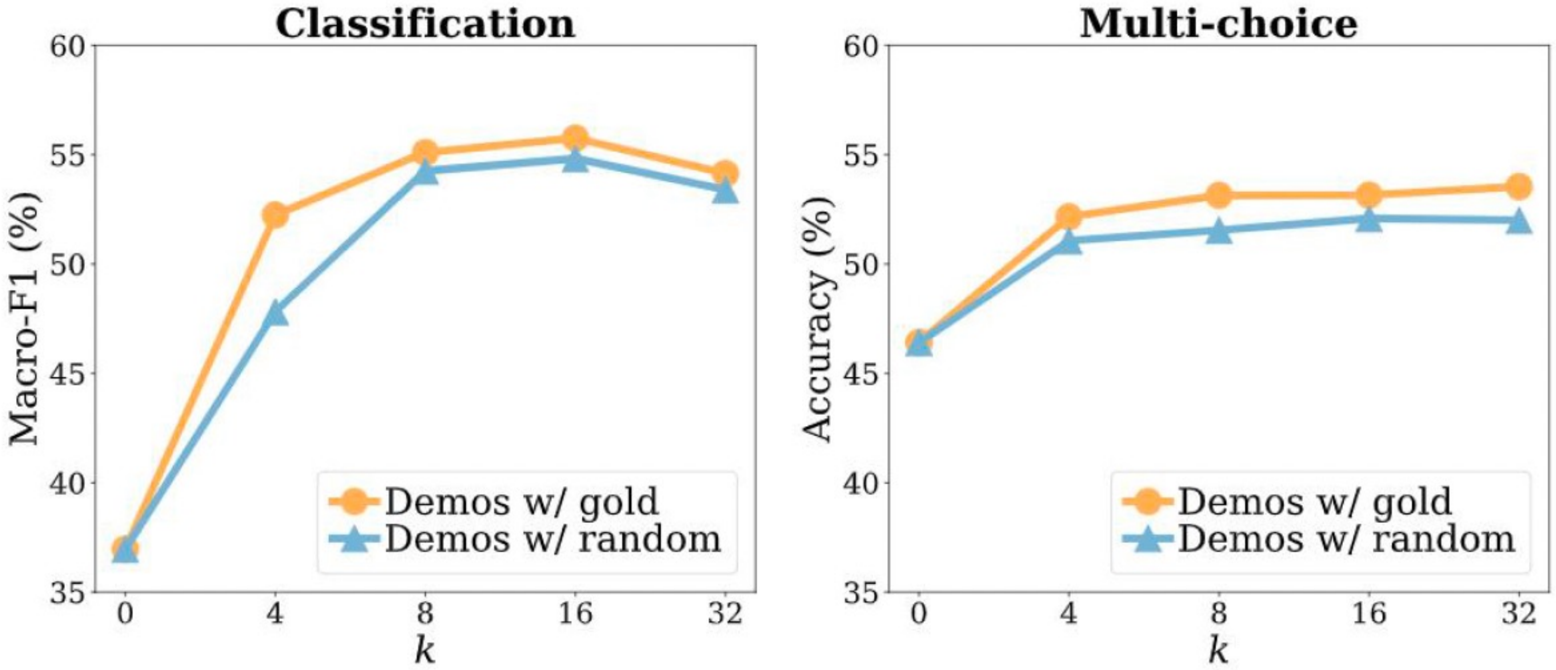


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

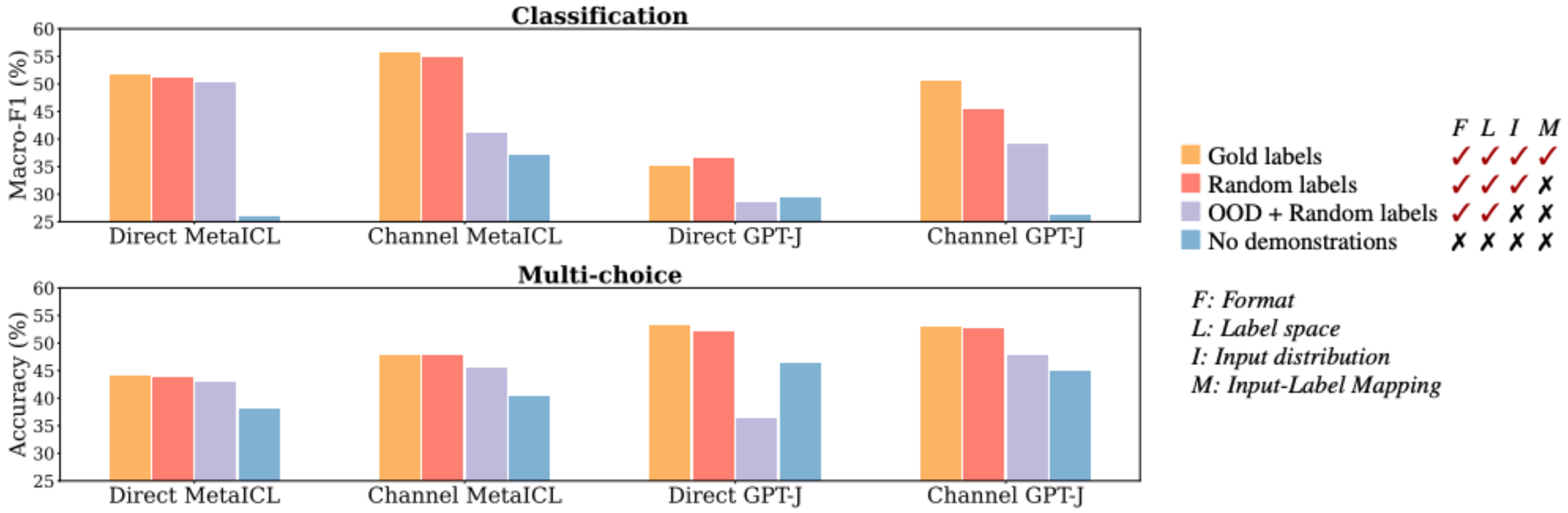
Experiment 3: Varying Numbers of Examples

- A small number of examples can already improve the performance
- Larger number of examples may result in performance convergence

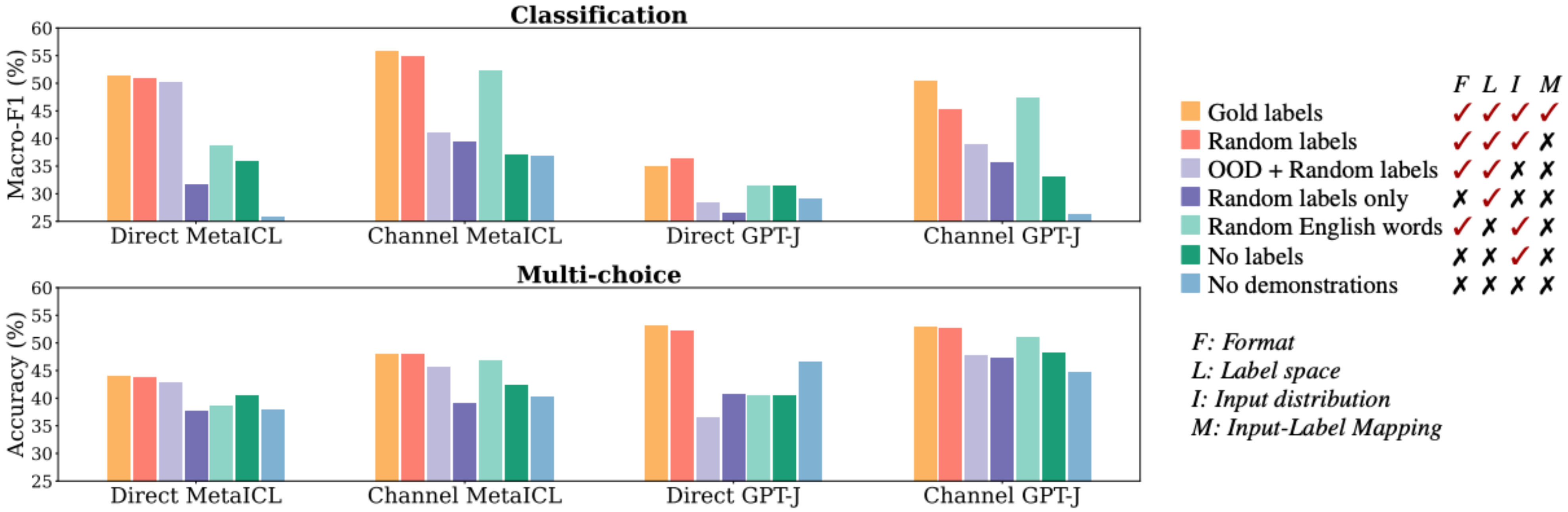


Experiment 4: Input Text Distribution

- Change the input example questions $x_1, x_2, \ldots, x_k$ to randomly sampled k sentences from external corpus, paired with random labels
- Significantly hurts the performance
- Model predicting texts conditioned on original input text is closer to the language modeling task



Experiment 4: Impact of the Input Format



- Observation: Keeping the format of input-label pairs is the key.

Content

- Scaling up Language Model: GPT-3
- Open-Source Model: Llama 2
- What Makes In-Context Learning Work?: Empirical Analysis
- **What Makes In-Context Learning Work?: Theoretical Analysis**

An Explanation of In-context Learning as Implicit Bayesian Inference

Sang Michael Xie
Stanford University
xie@cs.stanford.edu

Aditi Raghunathan
Stanford University
aditir@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

Tengyu Ma
Stanford University
tengyuma@cs.stanford.edu

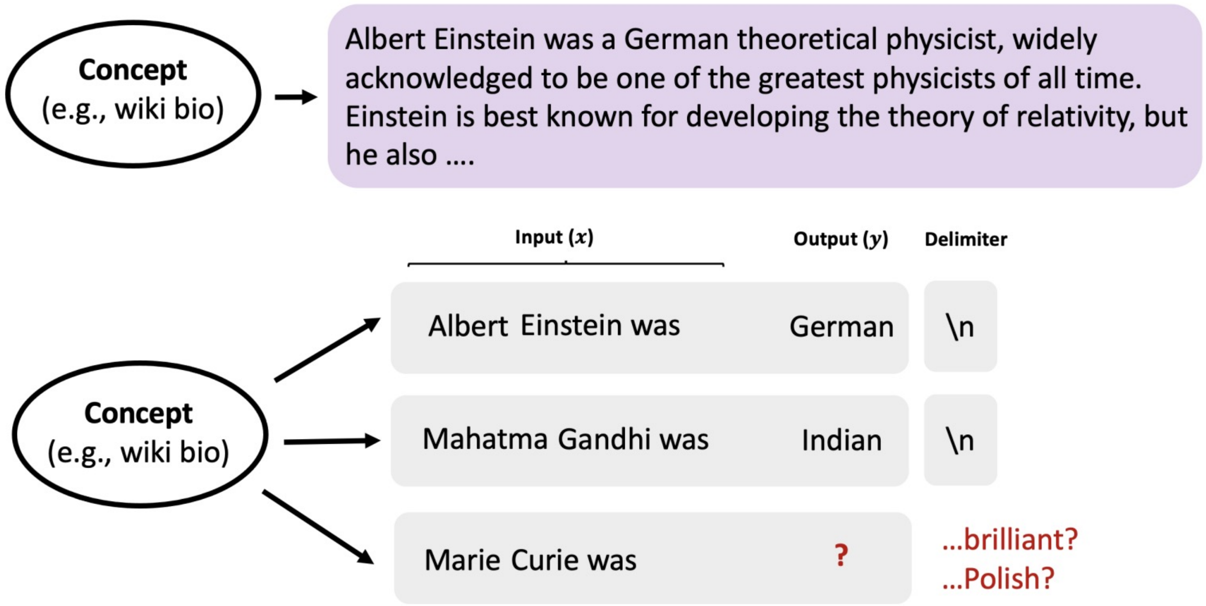
https://arxiv.org/pdf/2111.02080

Overview

- Proposes a Bayesian inference view for in-context learning with a mathematical proof
- Suggests that language models infer the concept for the current task before predicting the label

Mismatch between Pre-training and In-Context Learning

- Pre-training:
  - Next Token Prediction
- In-context learning:
  - Learn from examples
  - How does this work?



Text Prediction as Task Recognition

- Assumption: Language models are retrieving a learned concept to do in-context learning task
- What is a concept?
  - A latent variable θ that describes a distribution of words with semantic relations

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. // ________

LM

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. // ________

LM

$$p(\text{output}|\text{prompt}) = \int_{\text{concept}} p(\text{output}|\text{concept}, \text{prompt})p(\text{concept}|\text{prompt})d(\text{concept}).$$

Reformulating Inference

- Inferring answer y from examples S_n and question x_{test}

$$p(y|S_n, x_{\text{test}}) = \int_\theta p(y|S_n, x_{\text{test}}, \theta)p(\theta|S_n, x_{\text{test}})d\theta$$

$$\propto \int_\theta \sum_{h_{\text{test}}^{\text{start}} \in \mathcal{H}} p(y|x_{\text{test}}, h_{\text{test}}^{\text{start}}, \theta)p(h_{\text{test}}^{\text{start}}|S_n, x_{\text{test}}, \theta) \exp(n \cdot r_n(\theta))p(\theta)d\theta$$

the hidden state of the first token in x_{test}

- where

$$r_n(\theta) = \frac{1}{n} \log \frac{p(S_n, x_{\text{test}}|\theta)}{p(S_n, x_{\text{test}}|\theta^*)}$$

- and

$$\lim_{n\to\infty} \frac{p(S_n, x_{\text{test}}|\theta)}{p(S_n, x_{\text{test}}|\theta^*)} = \lim_{n\to\infty} \exp(n \cdot r_n(\theta)) = 0 \text{ for } \theta \neq \theta^*$$

- θ^* is the shared prompt concept between n examples
- This indicates that language model inference is equivalent to sampling from a superposition of tasks

Reformulating Inference (Cont'd)

- Proving $$\lim_{n\to\infty} e^{nr_n(\theta)} = \lim_{n\to\infty} \frac{p(S_n, x_{test}|\theta)}{p(S_n, x_{test}|\theta^*)} = \mathbf{1}_{\theta^*}$$

- Generation the input sequence

$$[S_n, x_{\text{test}}] = [x_1, y_1, o^{\text{delim}}, x_2, y_2, o^{\text{delim}}, \dots, x_n, y_n, o^{\text{delim}}, x_{\text{test}}] \sim p_{\text{prompt}}$$

- Can be seen as generation of independent events $\mathrm{O_i} = [x_i, y_i, o^{delim}]$
- $p(S_n, x_{test}|\theta) \approx \prod_{i=1}^{n} p(O_i|\theta)$
- When context clues of all examples align, models make stronger assumptions about which task is being performed.

Summary

- Pre-trained Large Language Model
  - GPT-3
  - Llama 2
- What makes in-context learning work?
  - Empirical experiments
  - Theoretical analysis

Next Class

- (Multi-task) instruction tuning
- More training examples
- More complex tasks
- Train the model to be flexible to adapt to different kinds of task instructions