CSE 561A: Large Language Models

Fall 2025

Lecture 3: Scaling up Language Models and Their Emergent Abilities

Jiaxin Huang

Reminder

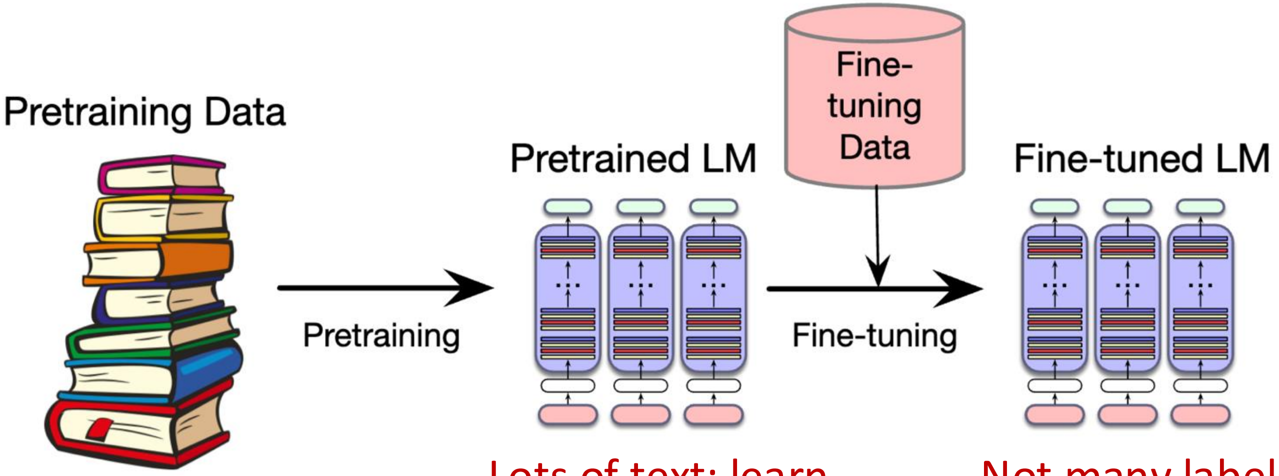
- For waitlist students: if it is your turn to get in, you will receive a notification/offer in **Workday (not in Email).** Please accept that that offer in 72 hours.
- The first student presentation lecture is on next Tuesday (Sept.9th)
- Presenters (on Sept.9th) please send your slides to me (cc the TAs) before Friday 12:00PM (Sept. 5th)
- First Assignment (preview question) will be due on Sept.8th

Content

- **Recap: Pre-training and Fine-tuning**
- Scaling up Language Models
- Emergent Abilities: In-context learning
- Open weight model version: The Llama series
- What Makes In-Context Learning Work?: Empirical Analysis
- Many-Shot In-Context Learning

Pretraining - Finetuning Paradigm

- Pretraining: trained with pretext tasks on large-scale text corpora
- Fine-tuning (continue training): adjust the pretrained model's parameters with fine-tuning data
- Fine-tuning data can have different forms:
    - Task-specific **labeled** data (e.g., sentiment classification, named entity recognition)
    - (Multi-turn) dialogue data (i.e., instruction tuning)

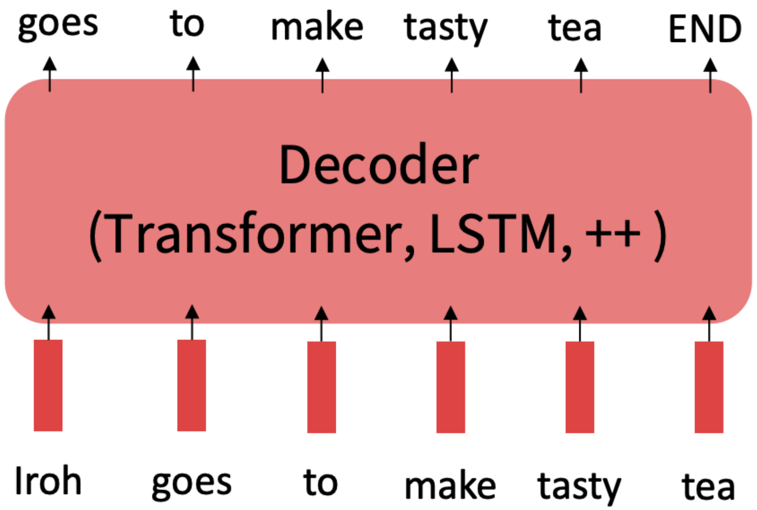


Lots of text; learn general things!

Not many labels; adapt to the task!

Decoder Pretraining (GPT)

- Decoder architecture is the prominent choice in large language models
- Pretraining decoders are first introduced in GPT (generative pretraining) models
- Recall the language modeling task: Model $p_\theta(w_t|w_{1:t-1})$, the probability distribution over words given their past contexts.
- Follow the standard language modeling (cross-entropy) objective

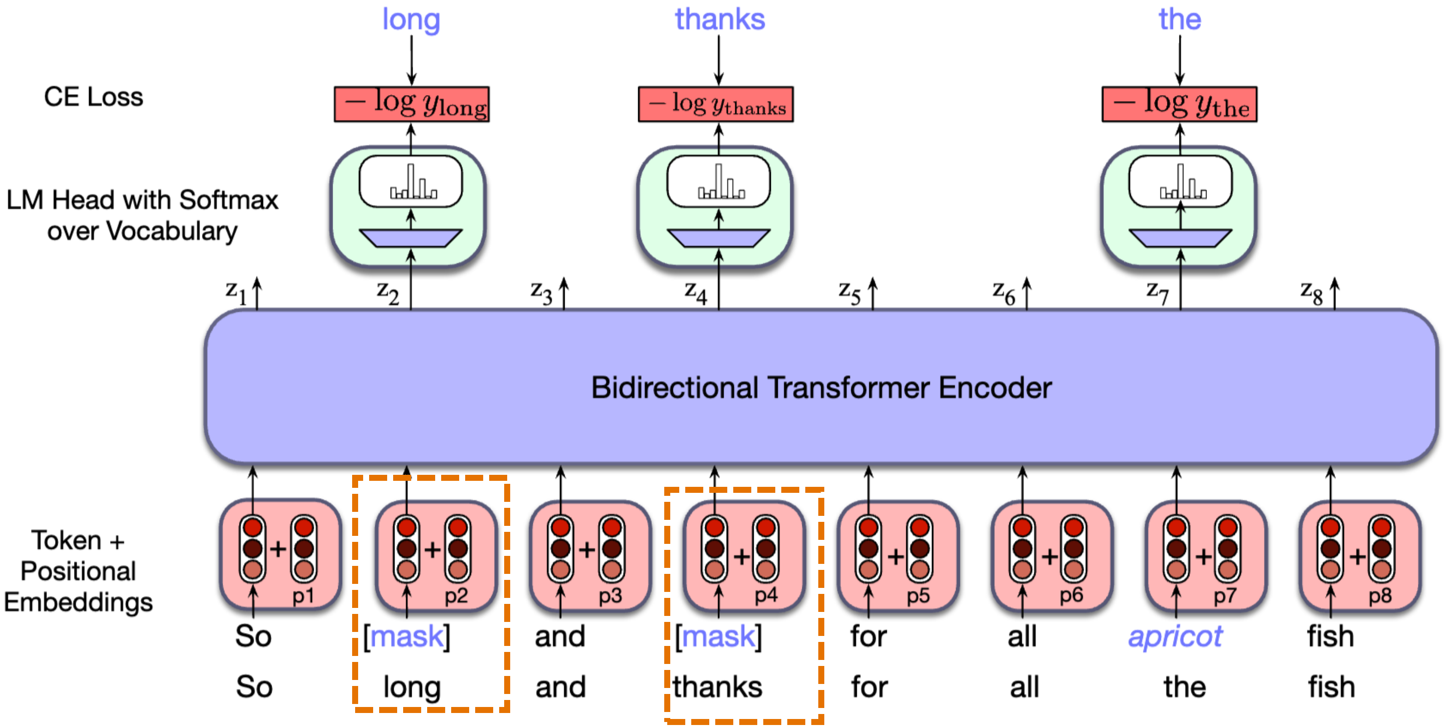


$$L = -\sum_{k=1}^{K} y_k \log(p_k)$$

[1] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI blog.

Encoder Pretraining: BERT

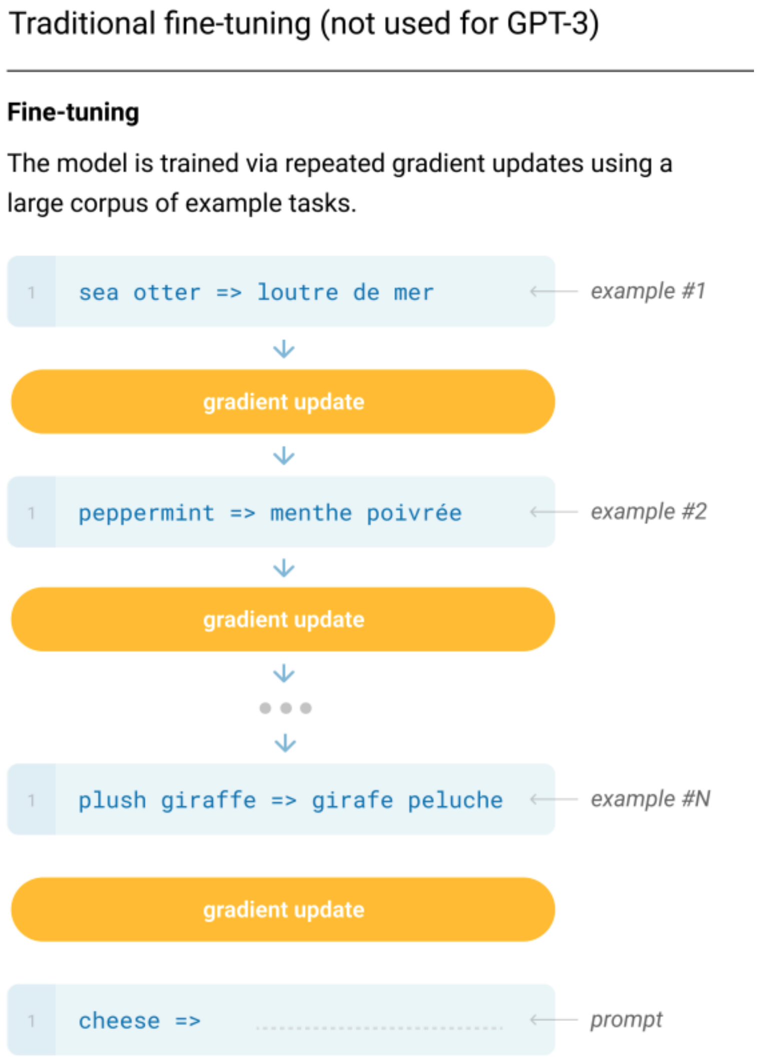
- BERT pretrains encoder models with bidirectionality
- **Masked language modeling** (MLM): With 15% words randomly masked or corrupted, the model learns bidirectional contextual information to predict the masked words



BERT: https://arxiv.org/pdf/1810.04805.pdf

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



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.

```
sea otter => loutre de mer
```

```
peppermint => menthe poivrée
```

```
plush giraffe => girafe peluche
```

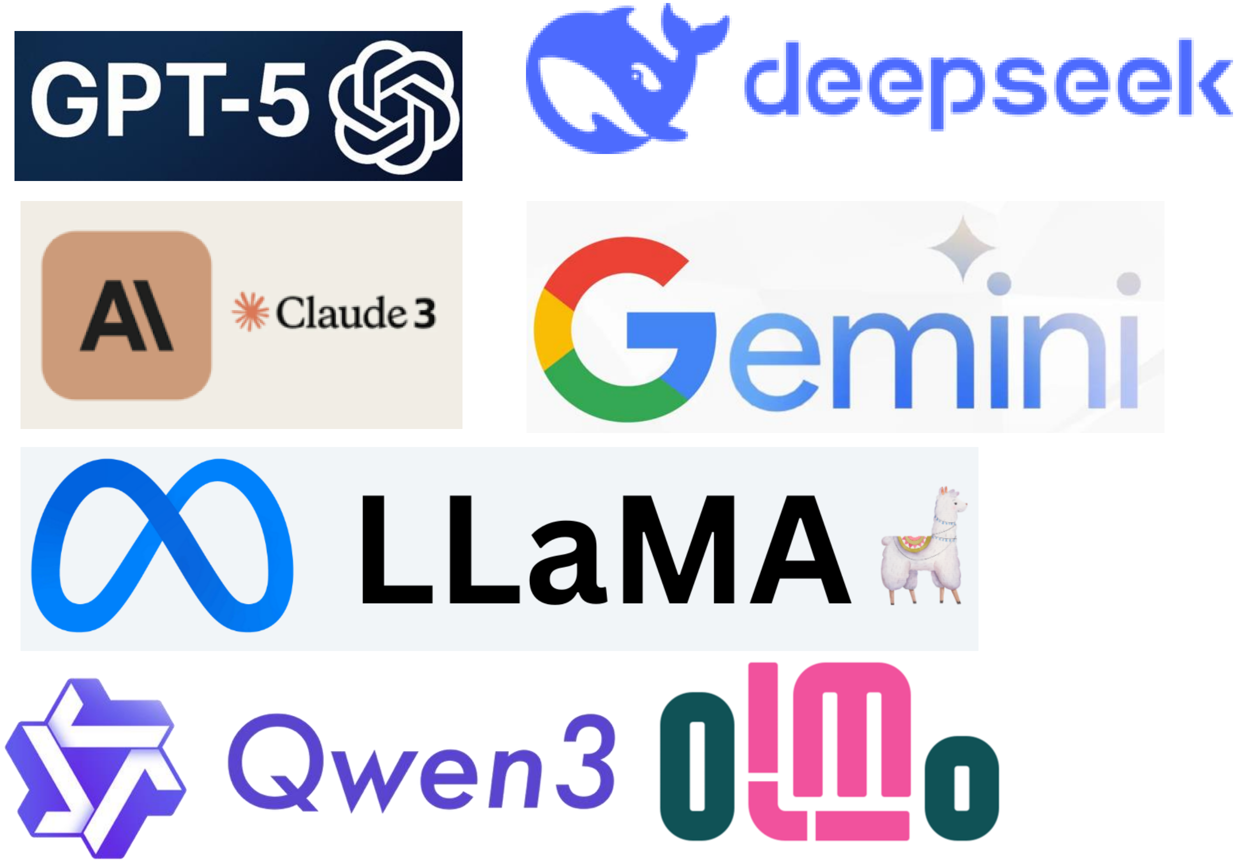
```
cheese =>
```

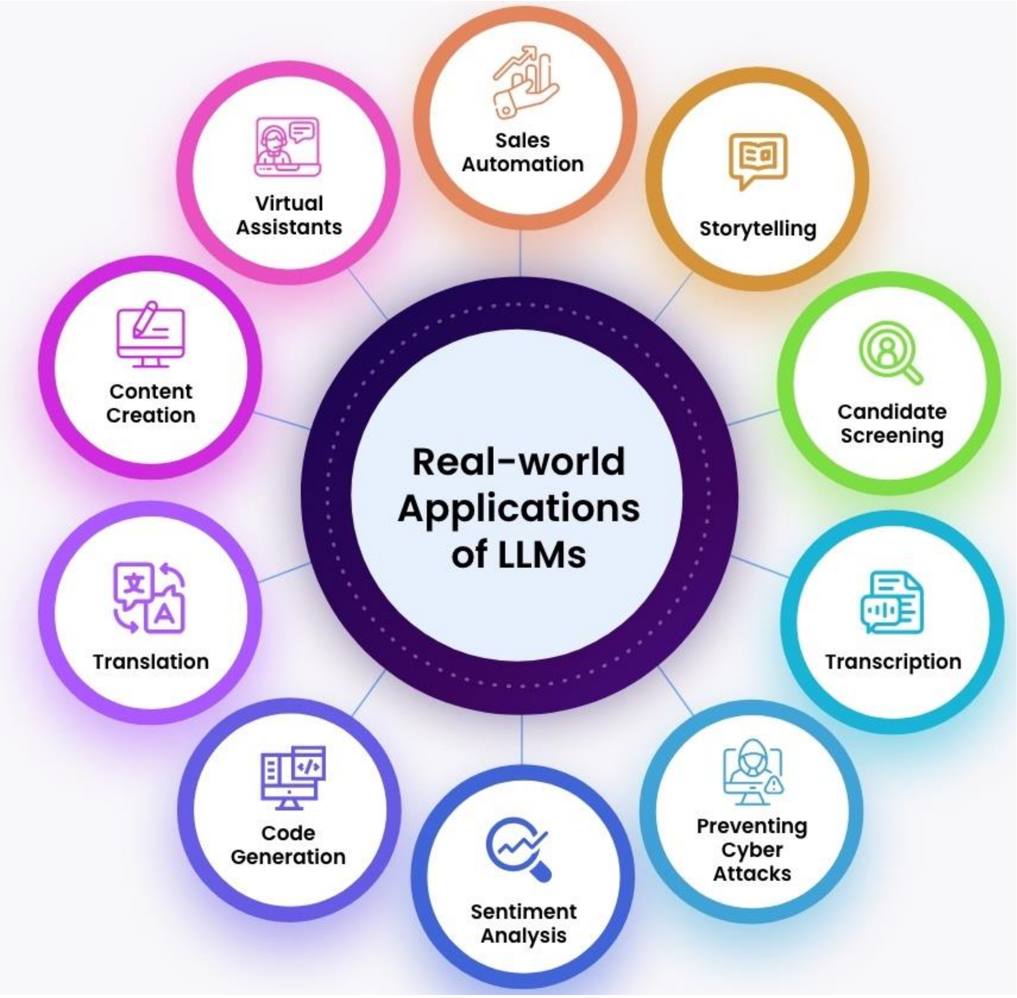
Content

- Recap: Pre-training and Fine-tuning
- **Scaling up Language Models**
- Emergent Abilities: In-context learning
- Open weight model version: The Llama series
- What Makes In-Context Learning Work?: Empirical Analysis
- Many-Shot In-Context Learning

A Plethora of Large Language Models

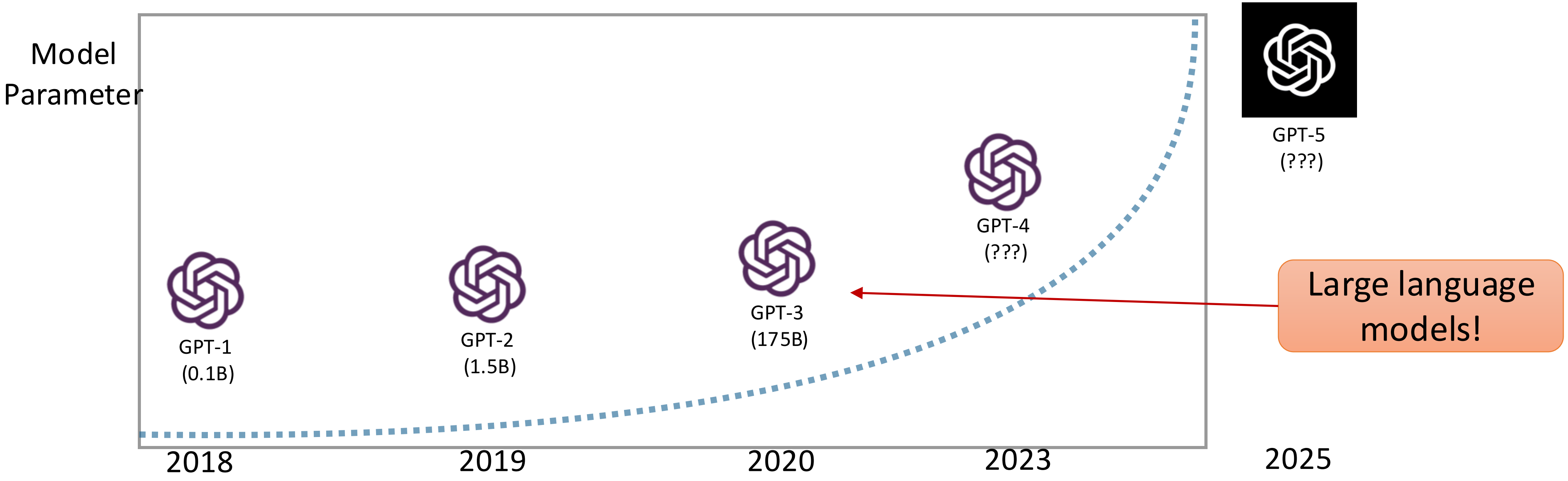
- Emerging large language models assisting our daily activities





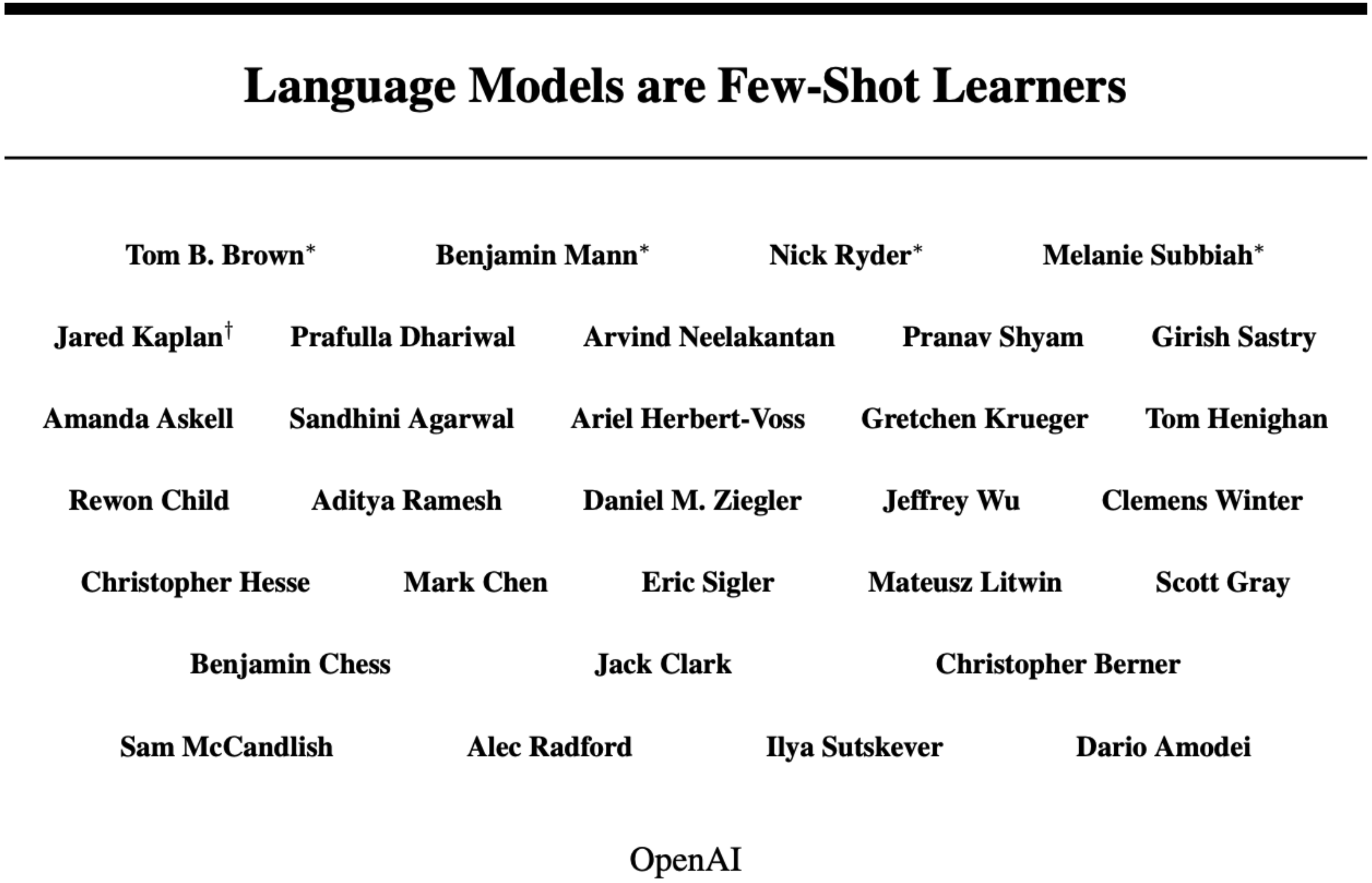
Next Class: Scaling up Language Models

- GPT-1 (2018): 12 layers, 117M parameters, trained in ~1 week
- GPT-2 (2019): 48 layers, 1.5B parameters, trained in ~1 month
- GPT-3 (2020): 96 layers, 175B parameters, trained in several months



Large language models!

Papers: (GPT-1) https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
(GPT-2) https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
(GPT-3) https://arxiv.org/pdf/2005.14165.pdf



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

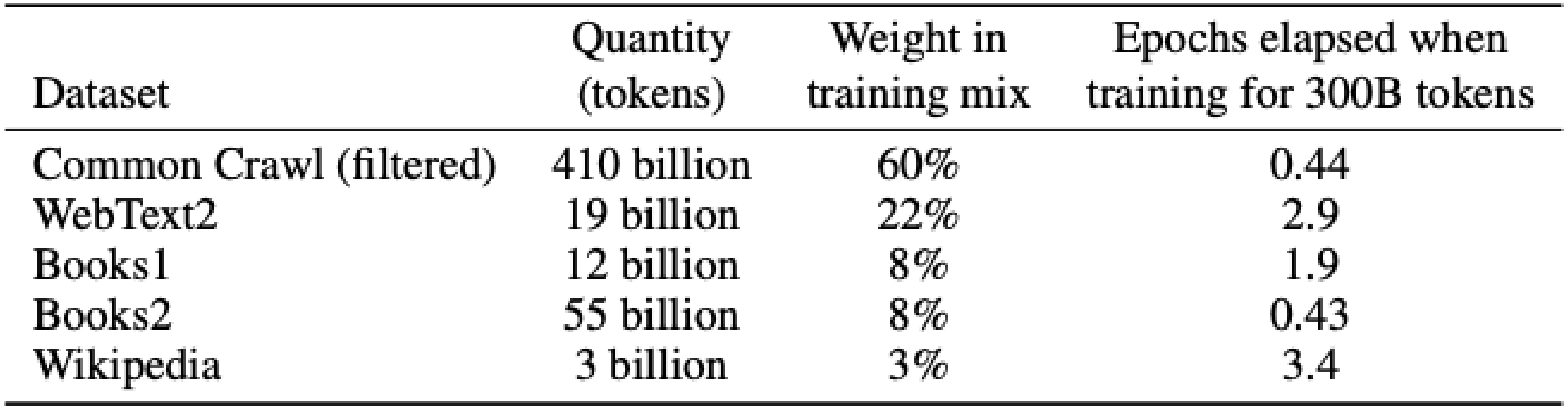
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

https://arxiv.org/pdf/2005.14165

Scaling up GPT Models – Pre-Training Data

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

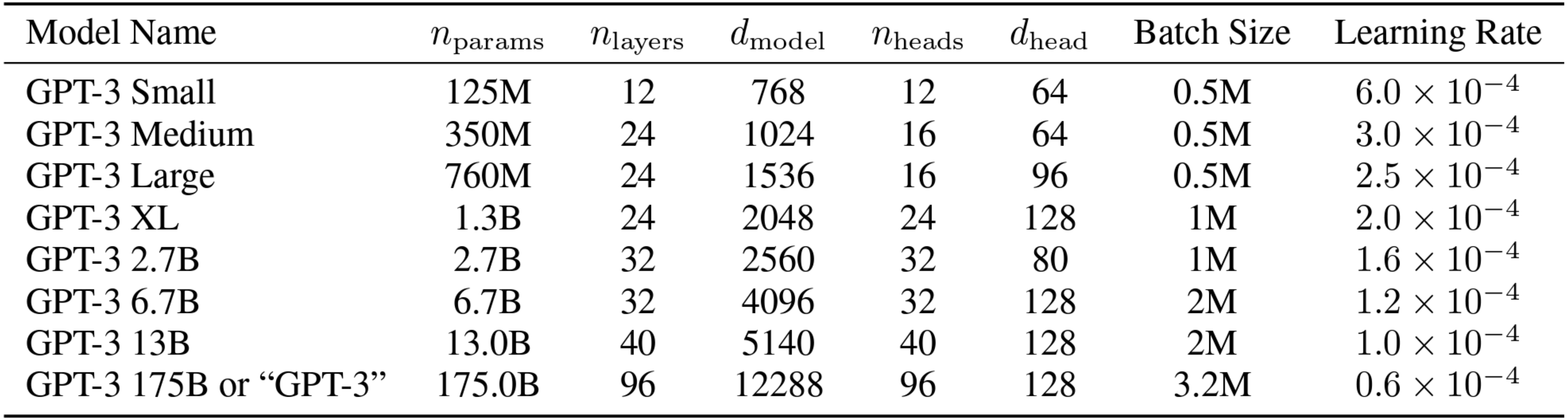


Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Training objective remains the same:

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

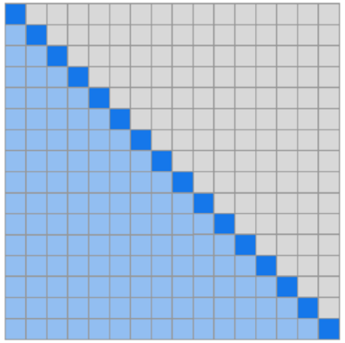
Scaling up GPT Models – Architecture



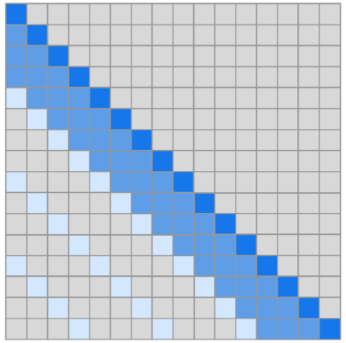
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	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

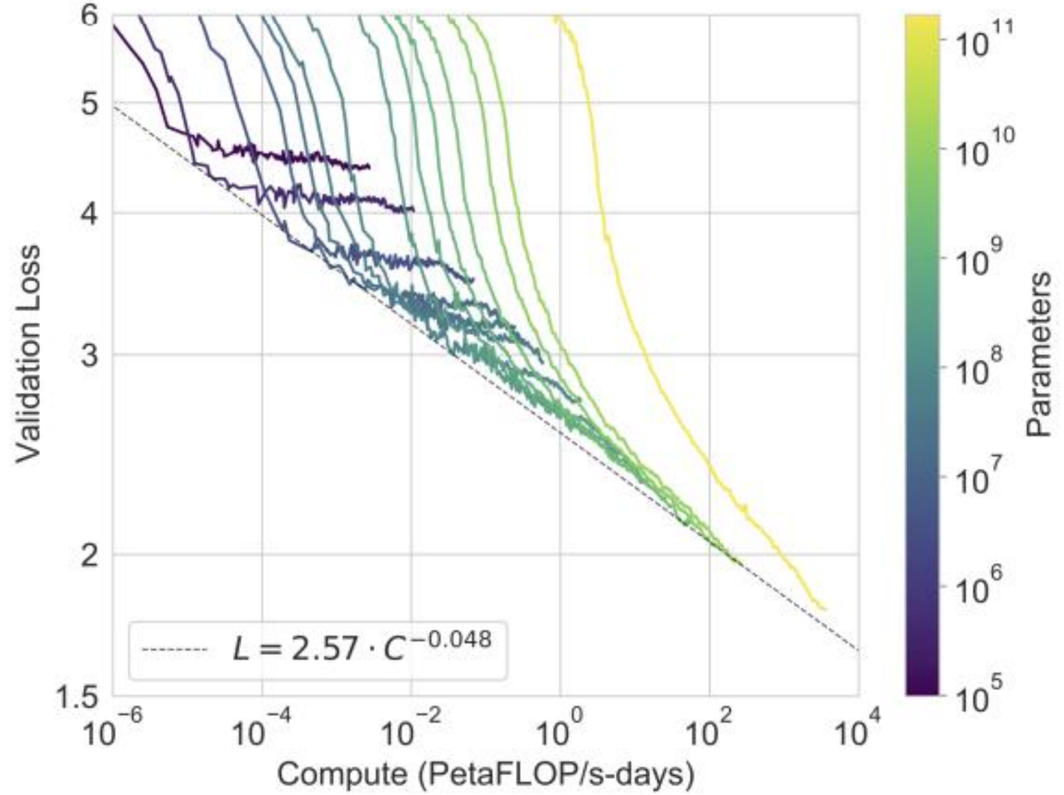


Sparse Attention: Tokens attend to sliding window

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

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



Content

- Recap: Pre-training and Fine-tuning
- Scaling up Language Models
- **Emergent Abilities: In-context learning**
- Open weight model version: The Llama series
- What Makes In-Context Learning Work?: Empirical Analysis
- Many-Shot In-Context Learning

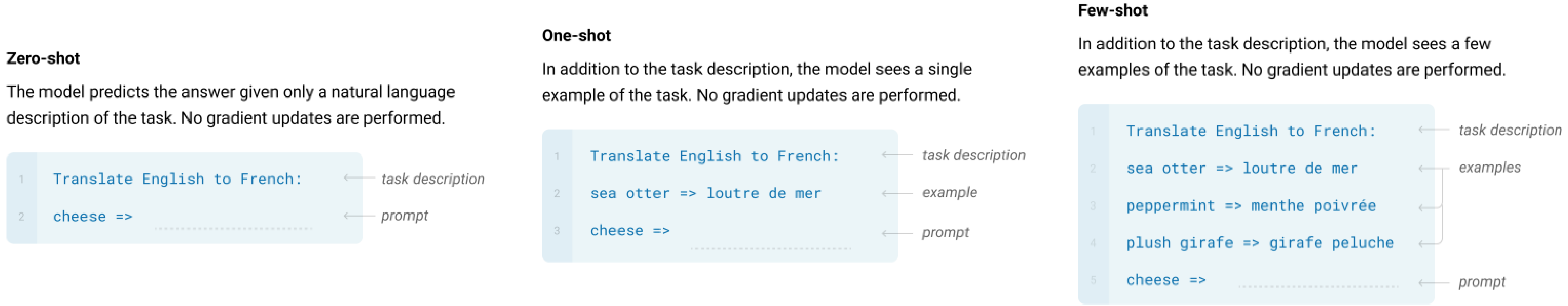
Emergent Ability

- Larger models develop **emergent abilities**
  - Skills or capabilities that were not explicitly learned but arise as a result of model capacity
  - Larger models demonstrate surprising abilities in challenging tasks even when they were not explicitly trained for them
- Emergent capabilities typically become noticeable only when the model size reaches a certain threshold (cannot be predicted by small model's performance)

Emergent Abilities of Large Language Models: https://arxiv.org/pdf/2206.07682

Emergent Ability: In-Context Learning

- In-context learning is a type of few-shot learning
  - User provides a few examples of input-output pairs in the prompt
  - The model uses given examples to predict the output for new, similar inputs
- First studied in the GPT-3 paper (Language Models are Few-Shot Learners: https://arxiv.org/pdf/2005.14165)
- No model parameter updates



Zero-shot

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

```
Translate English to French:
cheese =>
```

One-shot

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

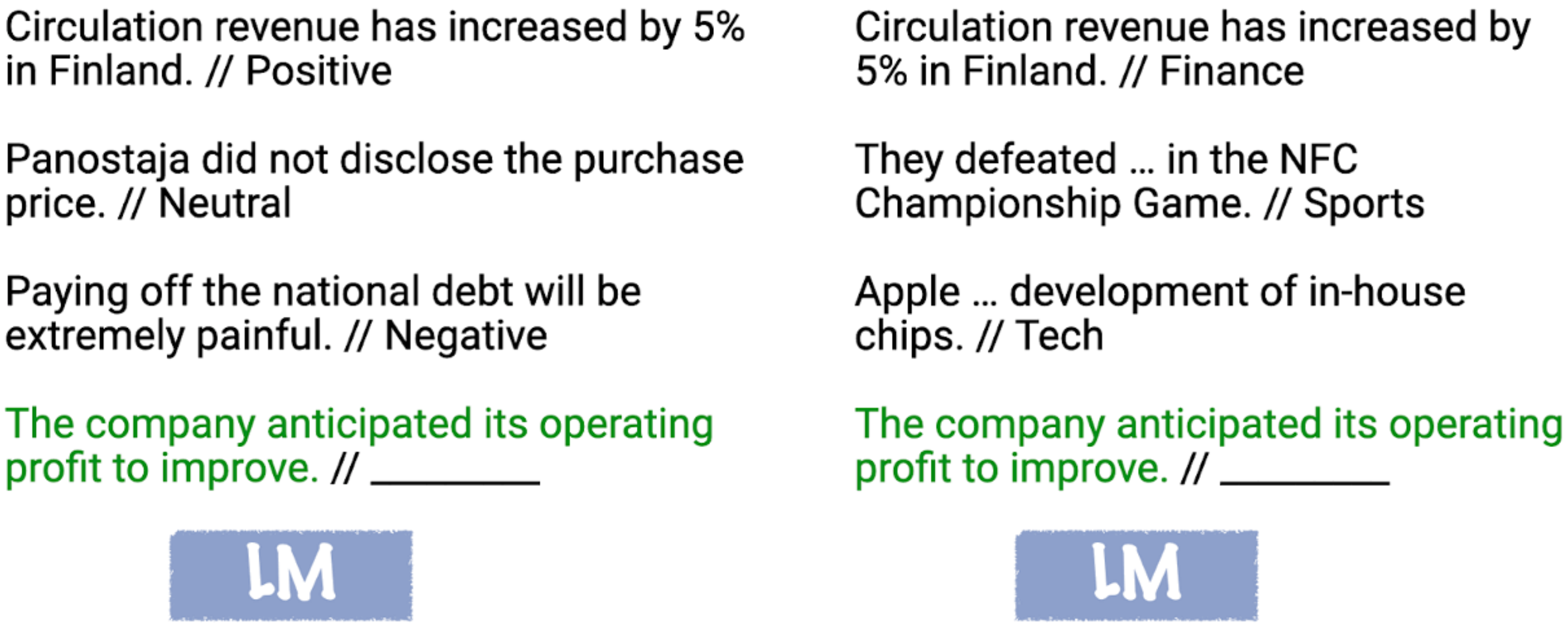
```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

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

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

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. // ________

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

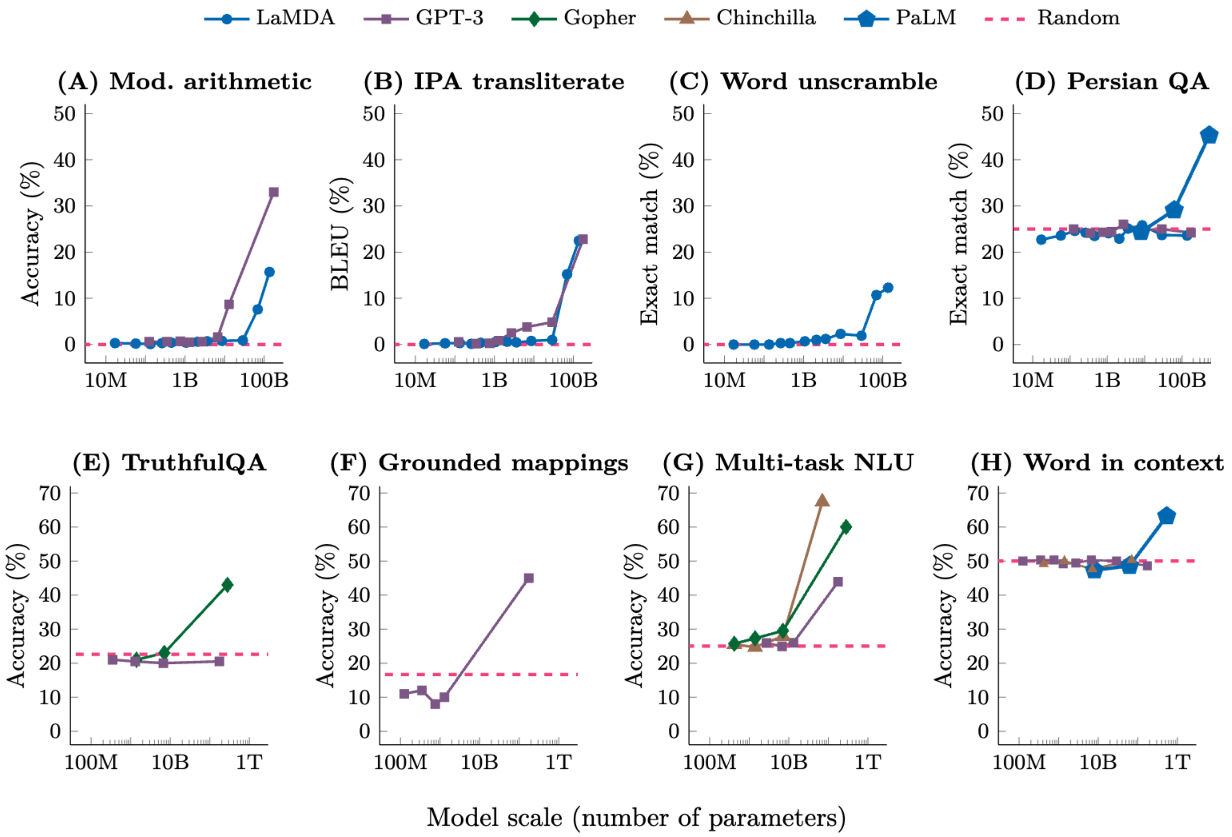
They defeated … in the NFC Championship Game. // Sports

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

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

Performance vs. Model Scale

- Models exhibit random performance until a certain scale, after which performance significantly increases



Tasks:
Arithmetic: addition, subtraction, multiplication
Transliteration
Recover a word from its scrambled letters
Persian question answering
Question answering (truthfully)
Grounded conceptual mappings
Multi-task understanding (math, history, law, ...)
Contextualized semantic understanding

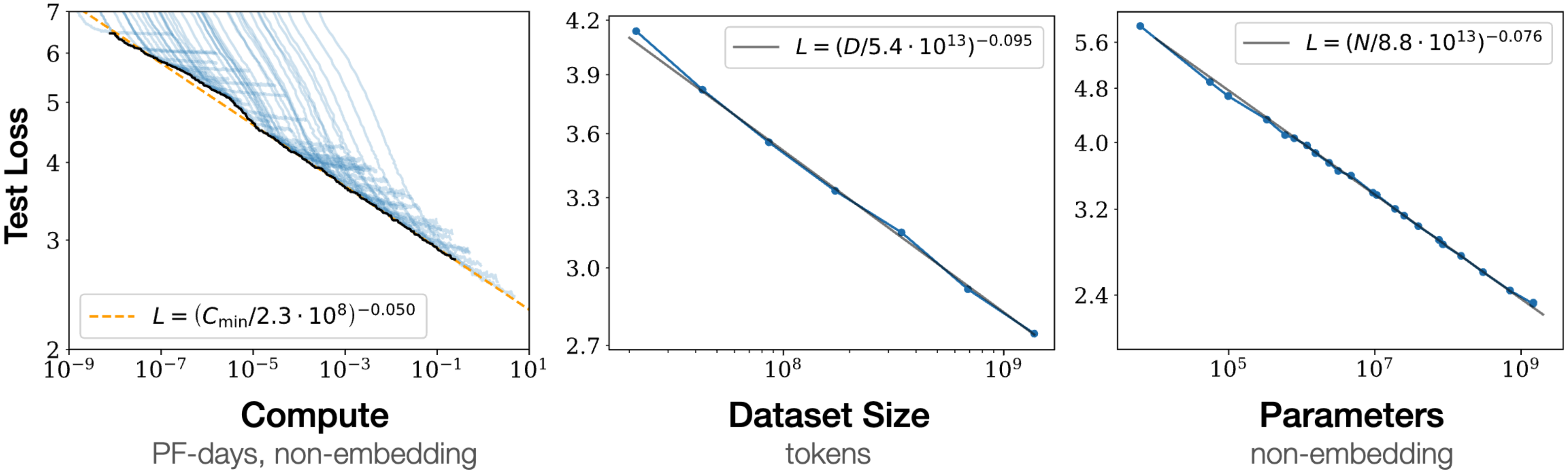
Figure source: https://arxiv.org/pdf/2206.07682

Scaling Laws of LLMs

- (Pretrained) LLM performance is mainly determined by 3 factors
  - Model size: the number of parameters
  - Dataset size: the amount of training data
  - Compute: the amount of floating point operations (FLOPs) used for training
- Scaling up LLMs involves scaling up the 3 factors
  - Add more parameters (adding more layers or having more model dimensions or both)
  - Add more data
  - Train for more iterations
- **Scaling laws**: study the correlation between the cross-entropy language modeling loss and the above three factors

Scaling Laws of LLMs

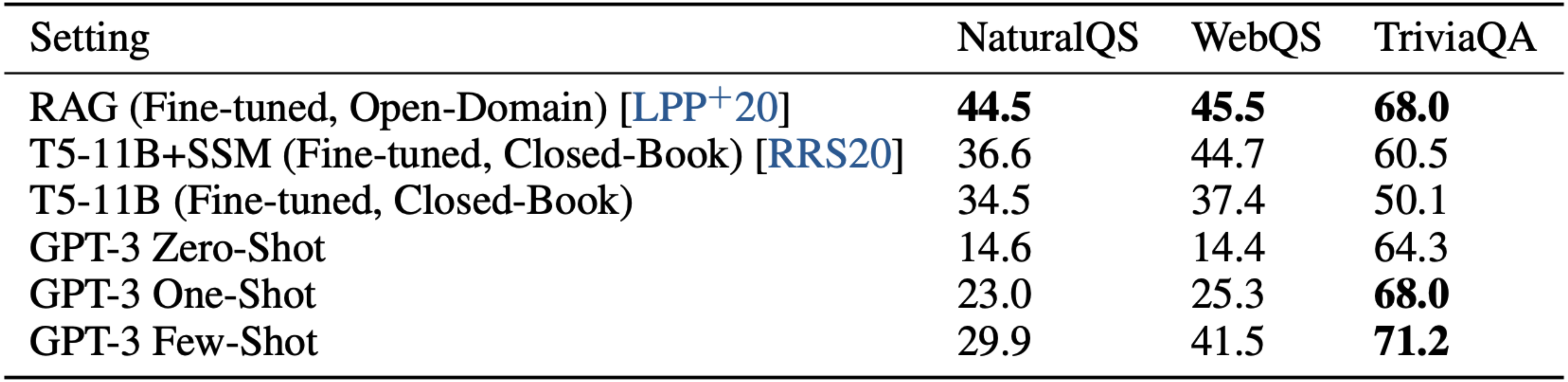
- Performance has a power-law relationship with each of the three scale factors (model size, dataset size, compute) when not bottlenecked by the other two



Paper: https://arxiv.org/pdf/2001.08361

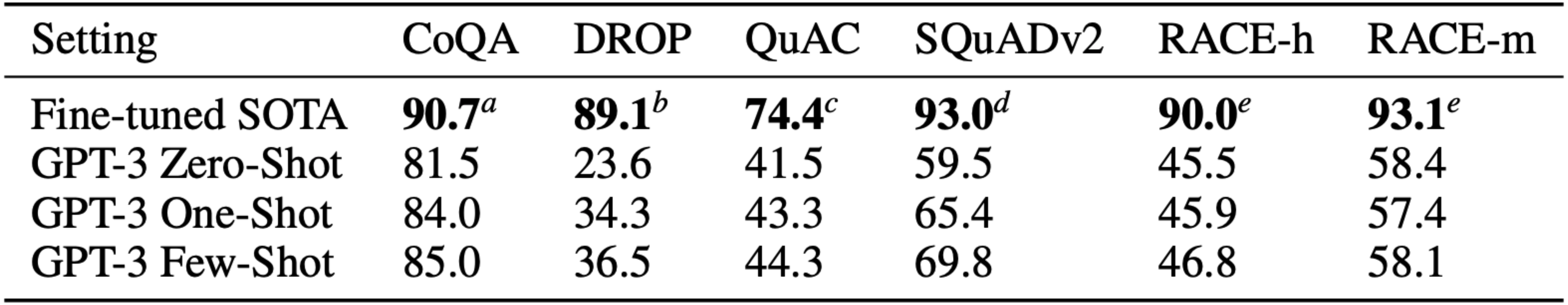
Evaluation on Question Answering Tasks

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



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. (We will discuss solutions to this next class!)

Limitations of GPT-3

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

Content

- Recap: Pre-training and Fine-tuning
- Scaling up Language Models
- Emergent Abilities: In-context learning
- **Open weight model version: The Llama series**
- What Makes In-Context Learning Work?: Empirical Analysis
- Many-Shot In-Context Learning

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

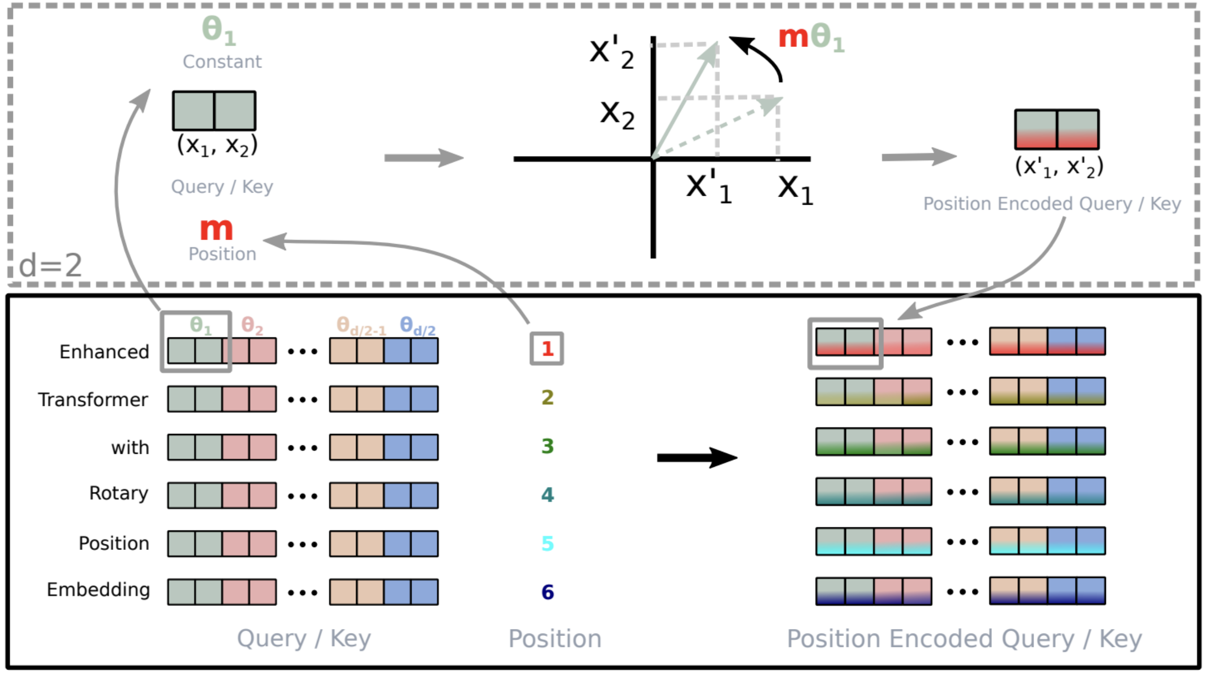
https://arxiv.org/pdf/2307.09288

Main Contribution

- Llama 2 was 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

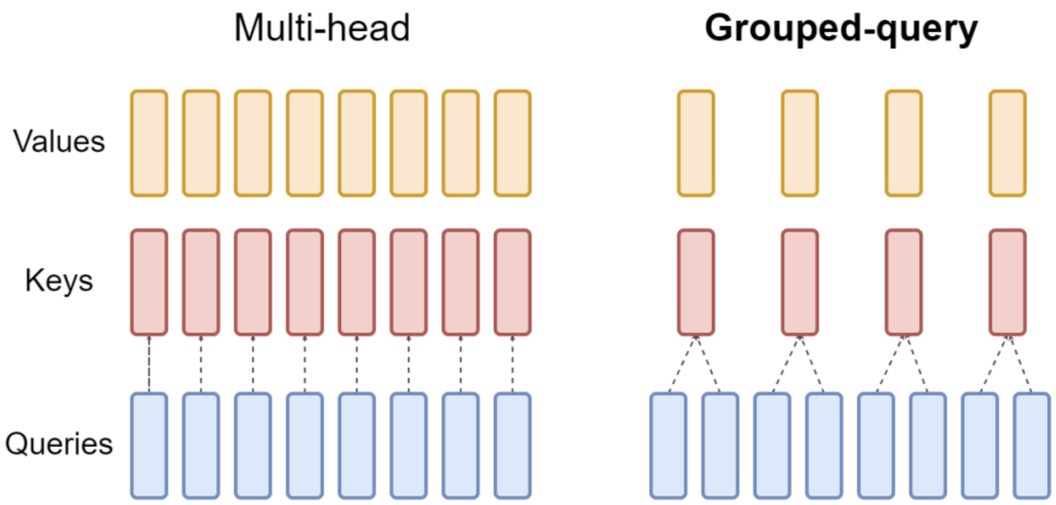
- Absolute positional encoding is simple, but may not generalize well in longer sequences
- Integrate relative position between tokens in the self-attention matrix



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

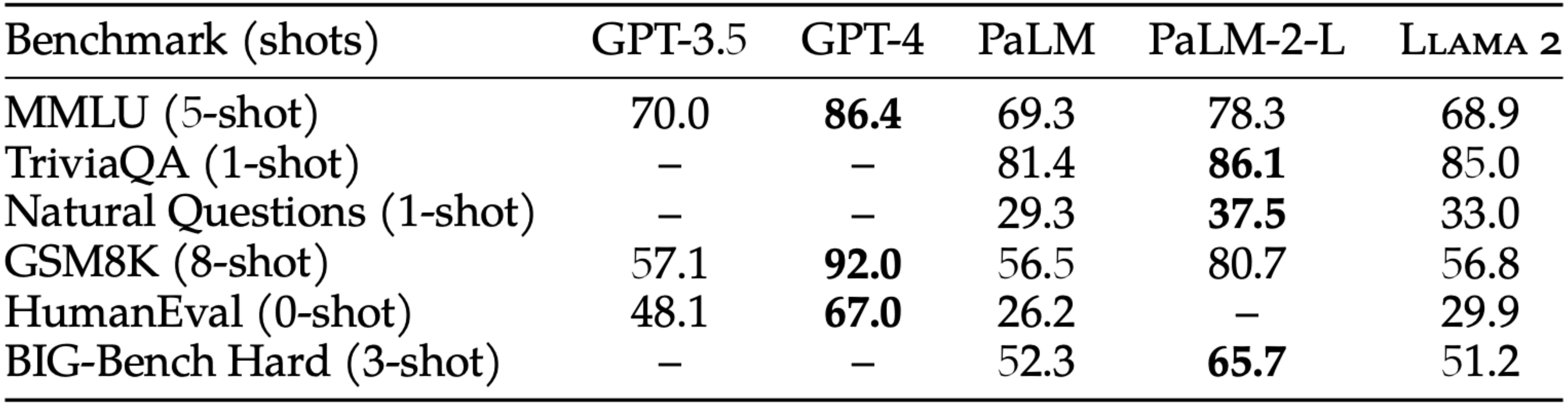
Llama 2 Improvement: Grouped-Query Attention

- 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

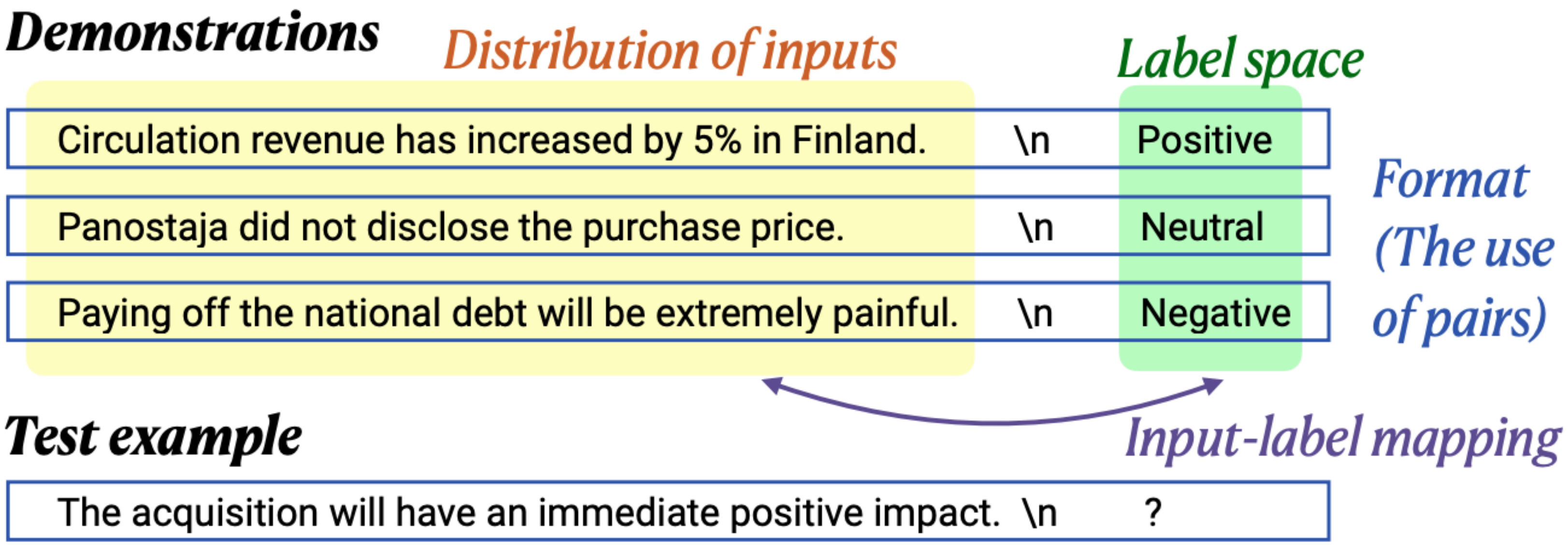
Content

- Recap: Pre-training and Fine-tuning
- Scaling up Language Models
- Emergent Abilities: In-context learning
- Open weight model version: The Llama series
- **What Makes In-Context Learning Work?: Empirical Analysis**
- Many-Shot In-Context Learning

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’)



Demonstrations

Circulation revenue has increased by 5% in Finland. \n Positive

Panostaja did not disclose the purchase price. \n Neutral

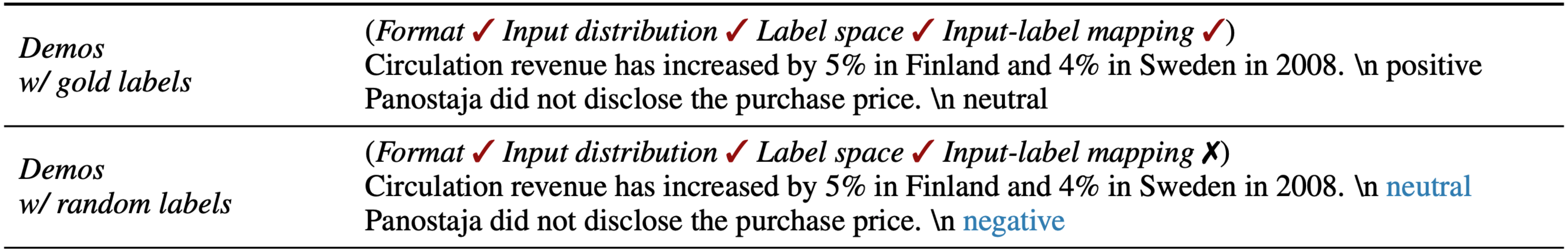
Paying off the national debt will be extremely painful. \n Negative

Test example

The acquisition will have an immediate positive impact. \n ?

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

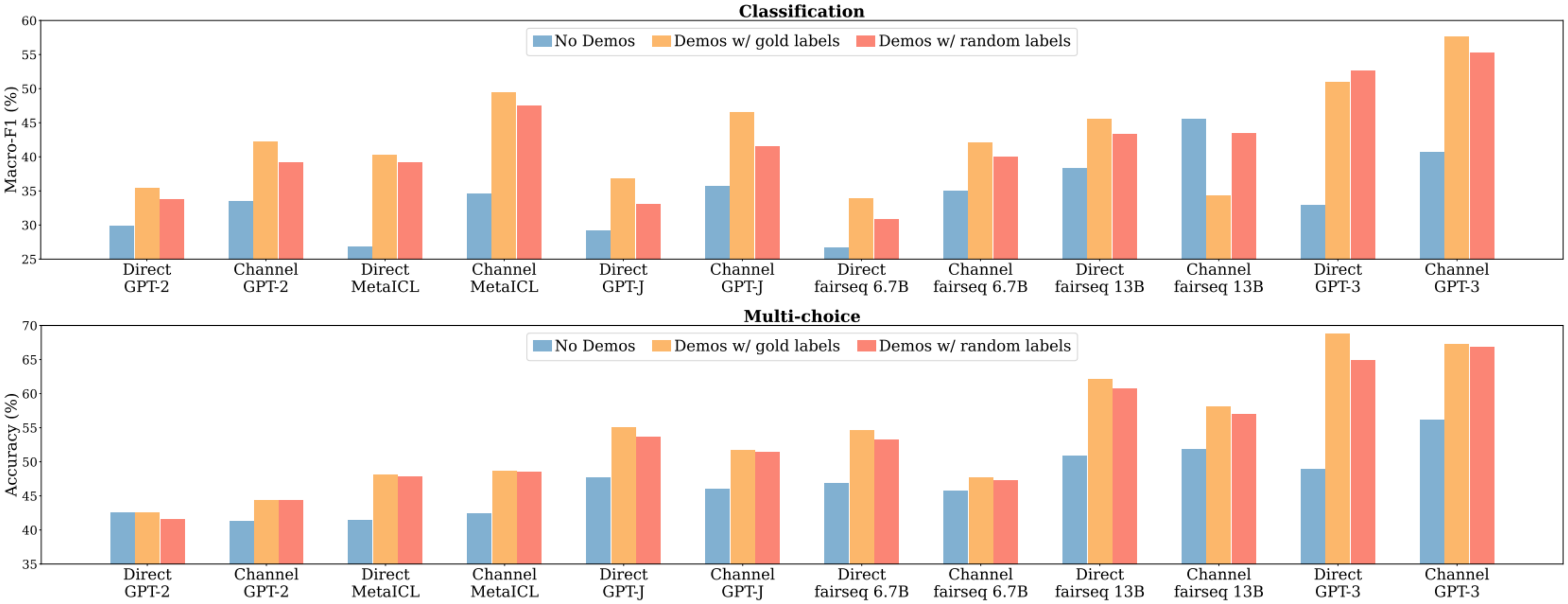
Experiment 1: Replace Gold Labels with Random Labels



Demos *w/ gold labels*	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✓) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n positive Panostaja did not disclose the purchase price. \n neutral
Demos *w/ random labels*	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n neutral Panostaja did not disclose the purchase price. \n negative

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

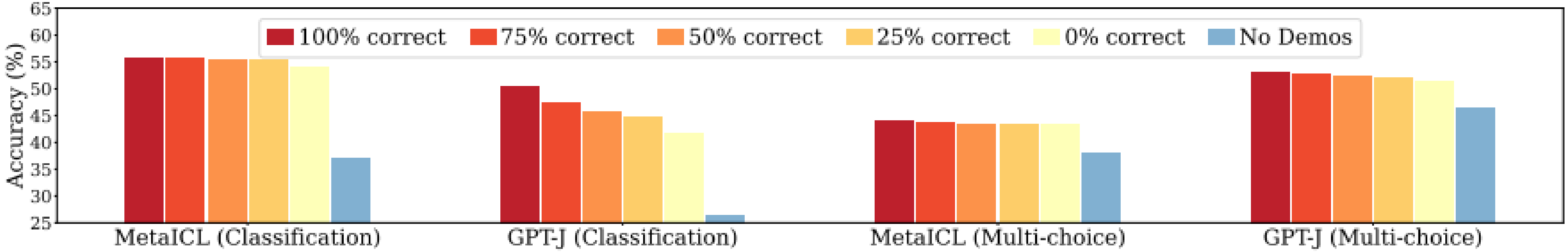
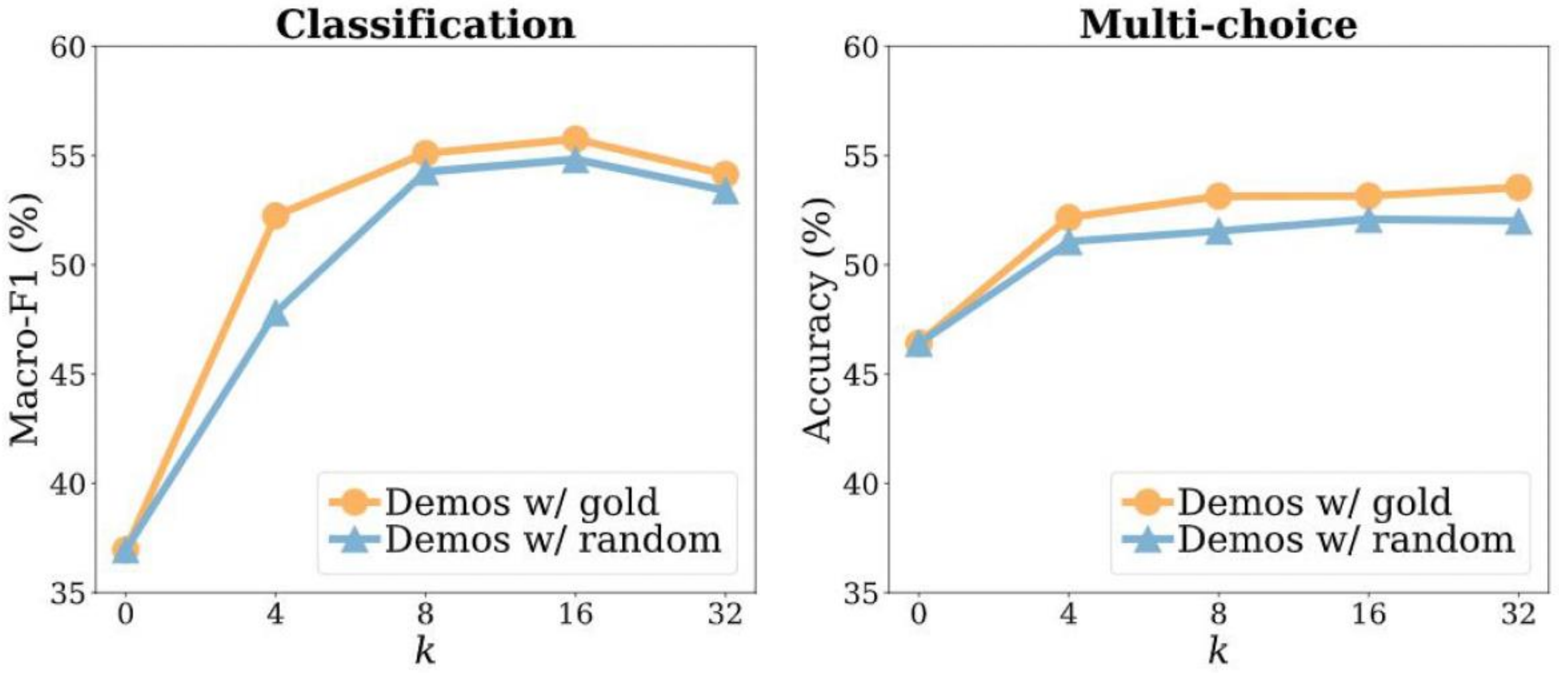


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

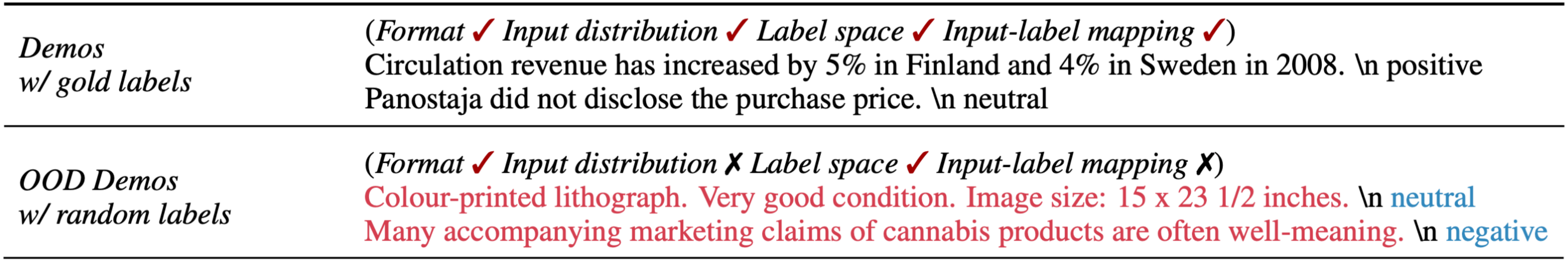
Experiment 3: Varying Numbers of Examples

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



Experiment 4: Input Text Distribution

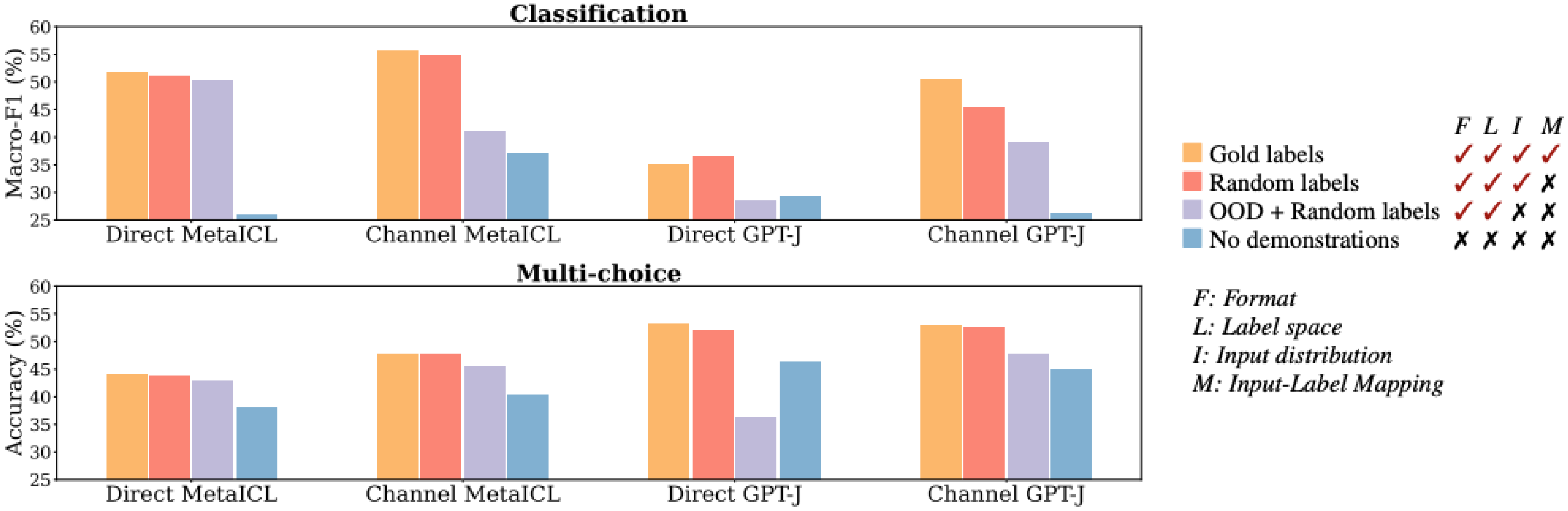
- Change the input example questions $x_1, x_2, \ldots, x_k$ to randomly sampled k sentences from external corpus, paired with random labels



Demos *w/ gold labels*	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✓) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n positive Panostaja did not disclose the purchase price. \n neutral
OOD Demos *w/ random labels*	(*Format* ✓ *Input distribution* ✗ *Label space* ✓ *Input-label mapping* ✗) Colour-printed lithograph. Very good condition. Image size: 15 x 23 1/2 inches. \n neutral Many accompanying marketing claims of cannabis products are often well-meaning. \n negative

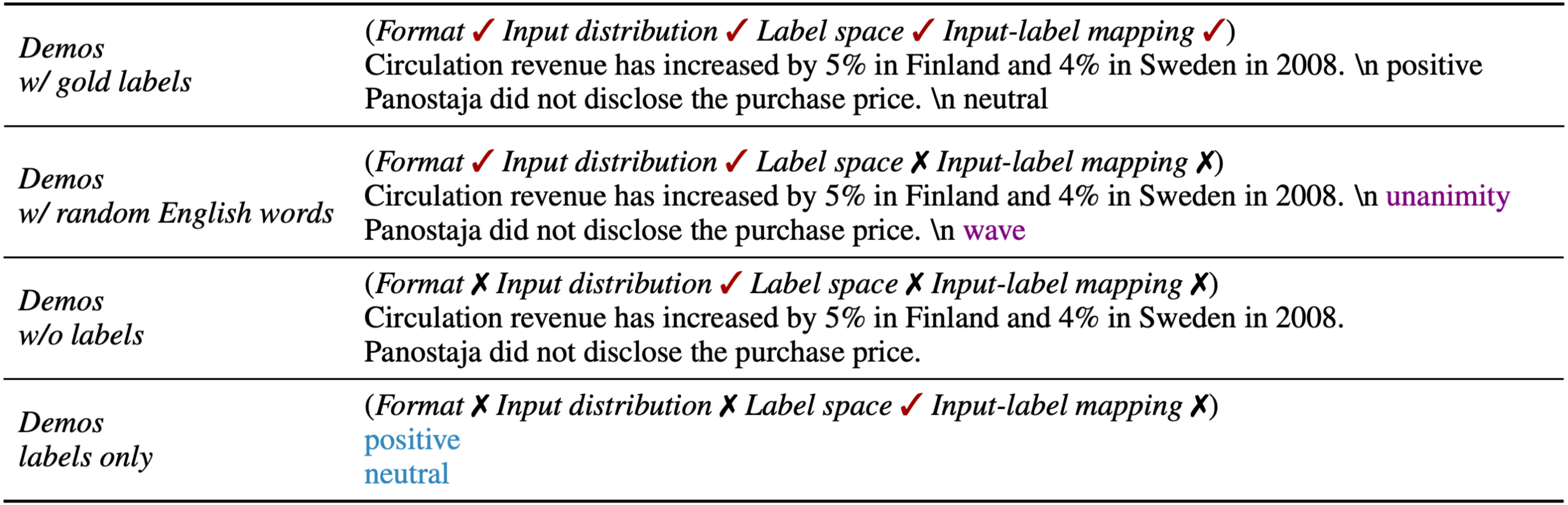
Experiment 4: Input Text Distribution

- 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



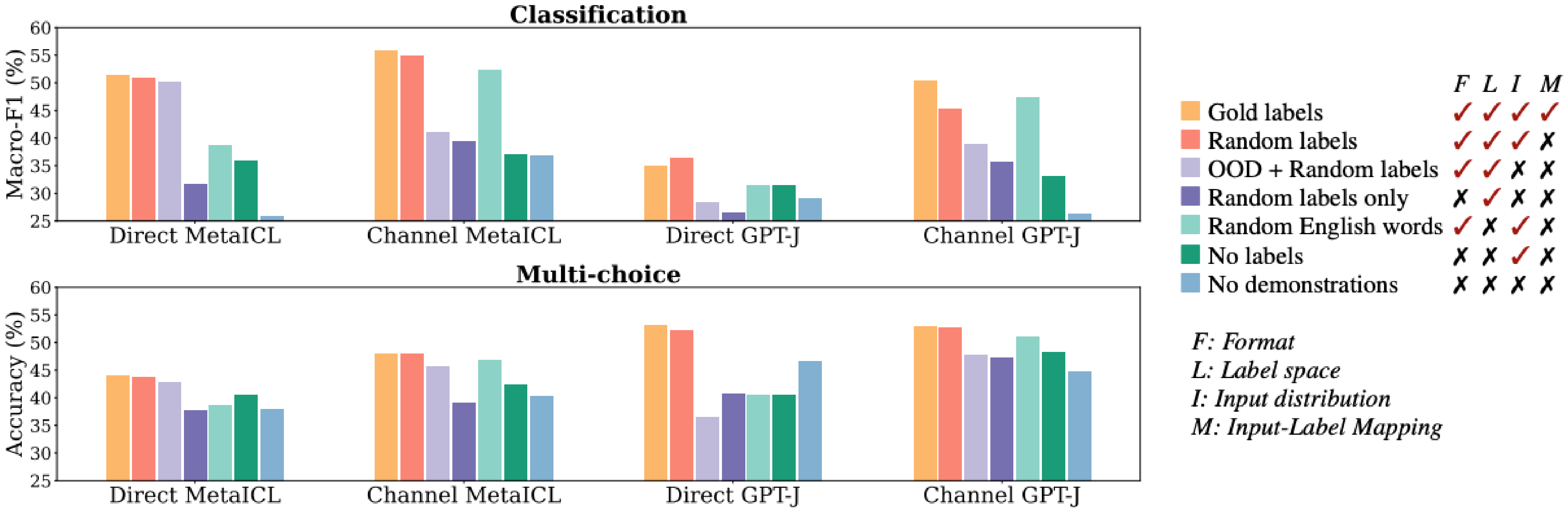
Experiment 5: Impact of the Input Format

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



Demos *w/ gold labels*	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✓) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n positive Panostaja did not disclose the purchase price. \n neutral
Demos *w/ random English words*	(*Format* ✓ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n unanimity Panostaja did not disclose the purchase price. \n wave
Demos *w/o labels*	(*Format* ✗ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. Panostaja did not disclose the purchase price.
Demos *labels only*	(*Format* ✗ *Input distribution* ✗ *Label space* ✓ *Input-label mapping* ✗) positive neutral

Experiment 5: Impact of the Input Format



	F	*L*	*I*	*M*
🟧 Gold labels	✓	✓	✓	✓
🟥 Random labels	✓	✓	✓	✗
🟪 OOD + Random labels	✓	✓	✗	✗
🟪 Random labels only	✗	✓	✗	✗
🟩 Random English words	✓	✗	✓	✗
🟩 No labels	✗	✗	✓	✗
🟦 No demonstrations	✗	✗	✗	✗

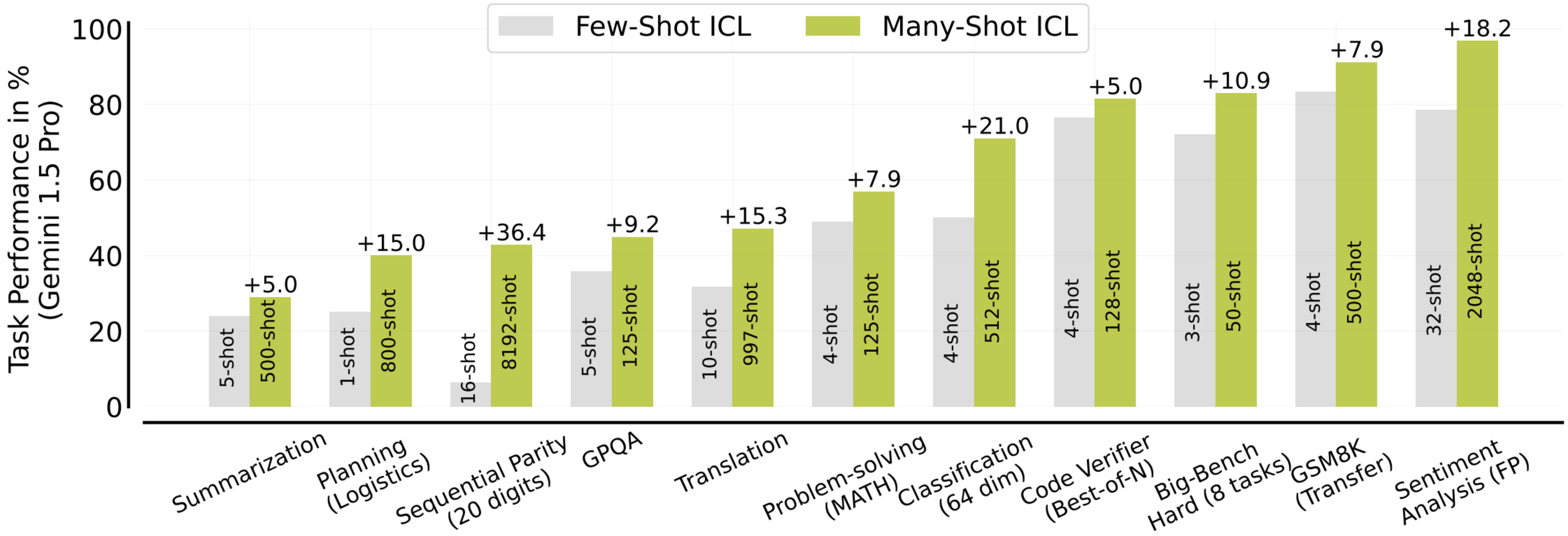
F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

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

Content

- Recap: Pre-training and Fine-tuning
- Scaling up Language Models
- Emergent Abilities: In-context learning
- Open weight model version: The Llama series
- What Makes In-Context Learning Work?: Empirical Analysis
- **Many-Shot In-Context Learning**

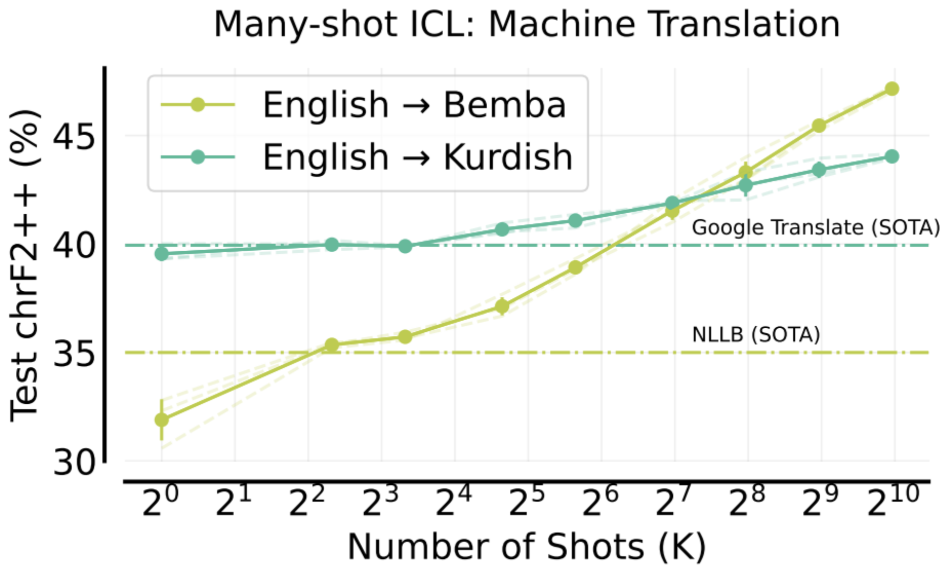
Performance with More Examples?

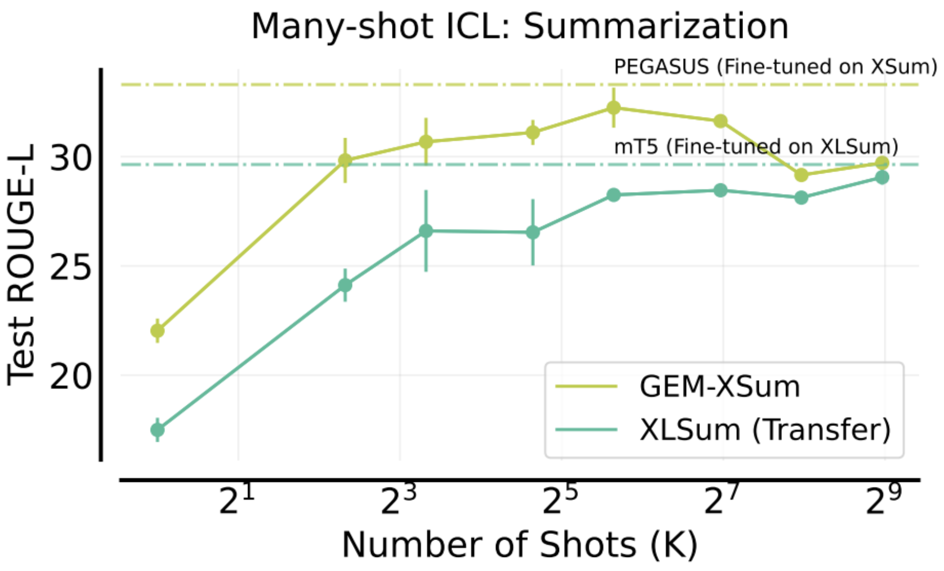


Many-Shot In-Context Learning. https://arxiv.org/pdf/2404.11018

How Many Examples are Enough?

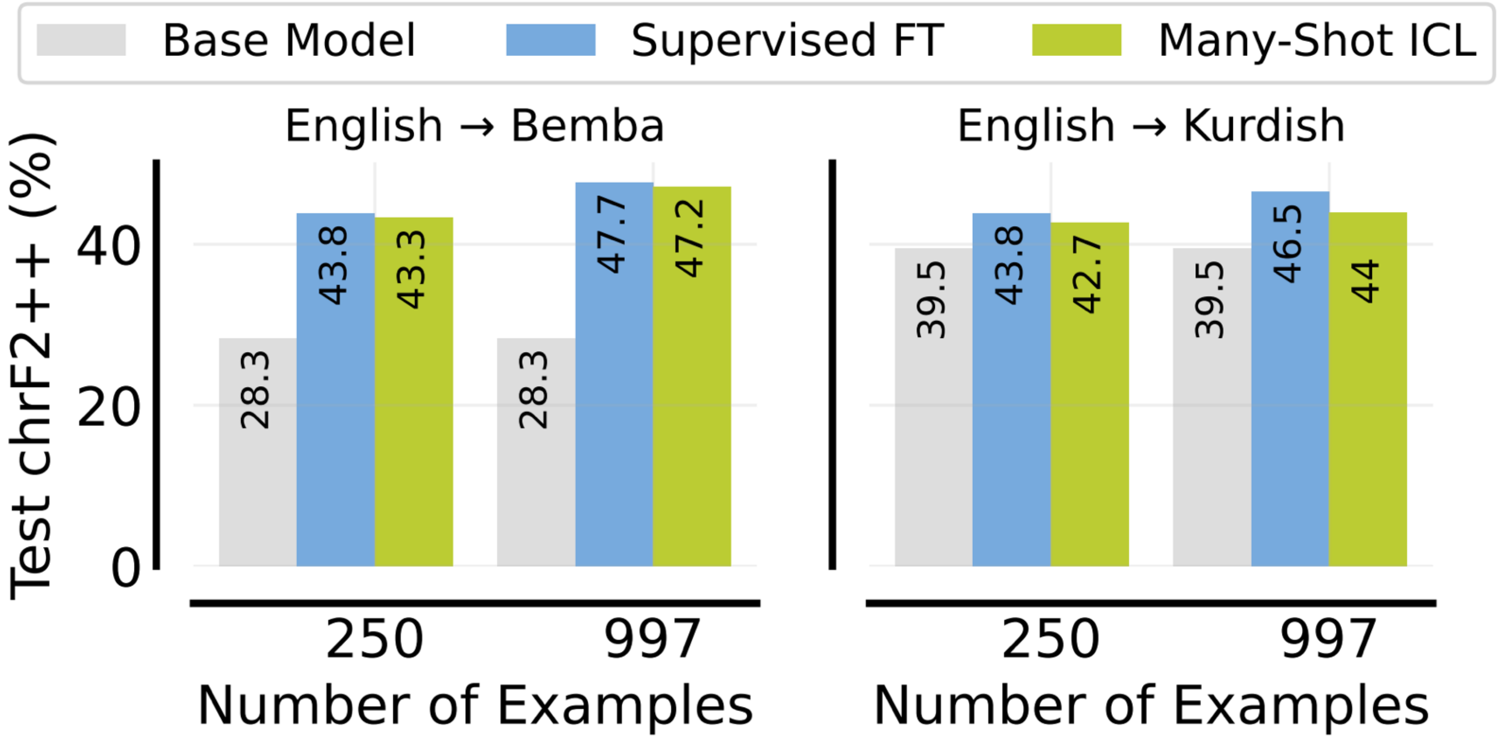
- The optimal number of examples varies across different tasks.


Many-shot ICL: Machine Translation


Many-shot ICL: Summarization

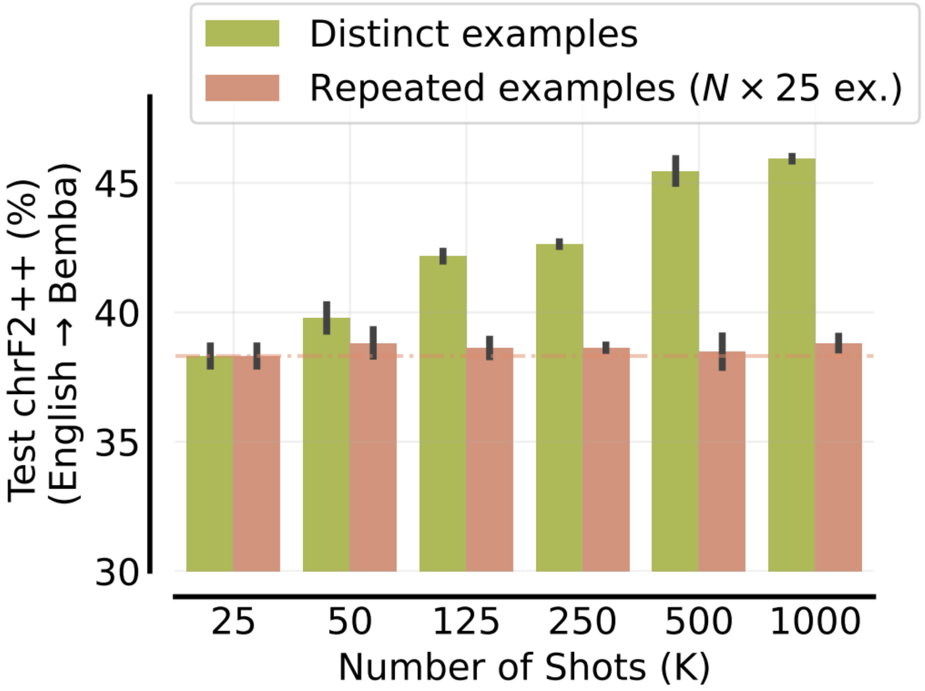
Many-Shot ICL vs. Supervised Fine-Tuning

- Supervised Fine-Tuning: larger training-time computation
- Many-Shot ICL: larger test-time computation



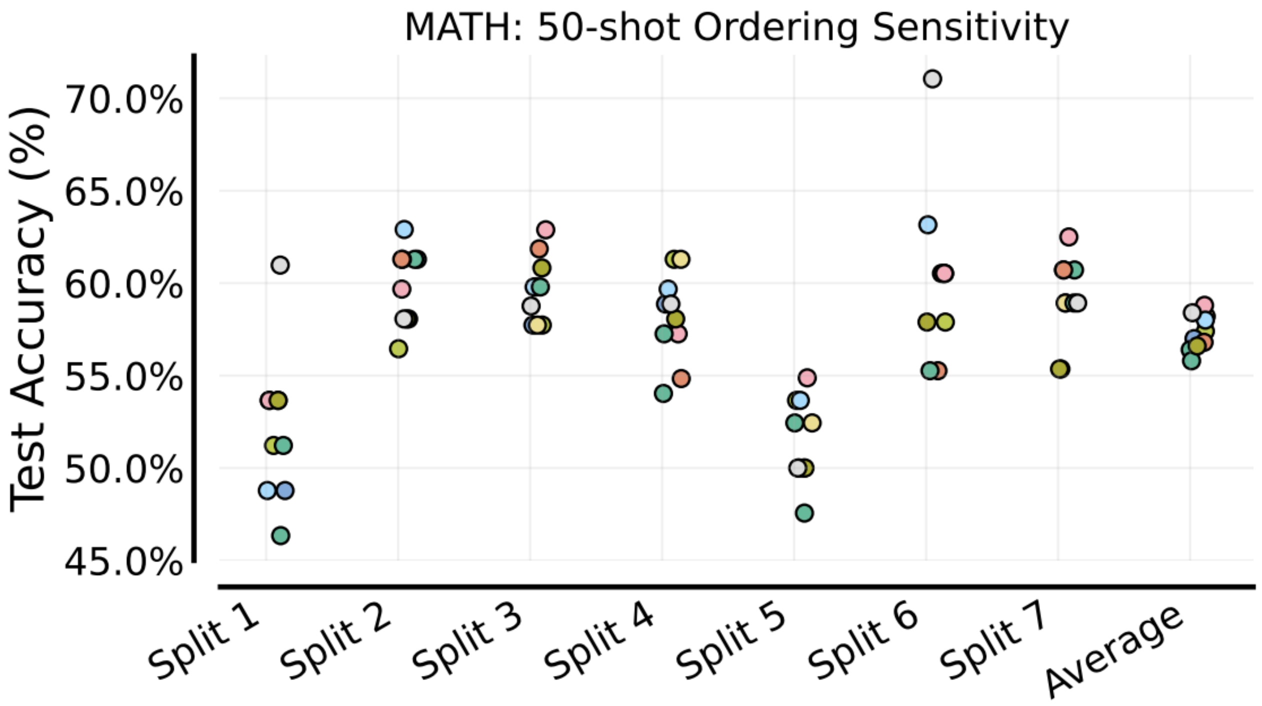
Does Repeated Examples Matter?

- With the same total number of examples, repeating examples could result in lower accuracy than using unique examples.
- Many-shot ICL mainly benefits from new information.



Does Re-Ordering Examples Matter?

- Each colored data point represents a different random ordering of 50 in-context examples provided to Gemini 1.5 Pro.
- Yes, the order significantly impacts model performance!

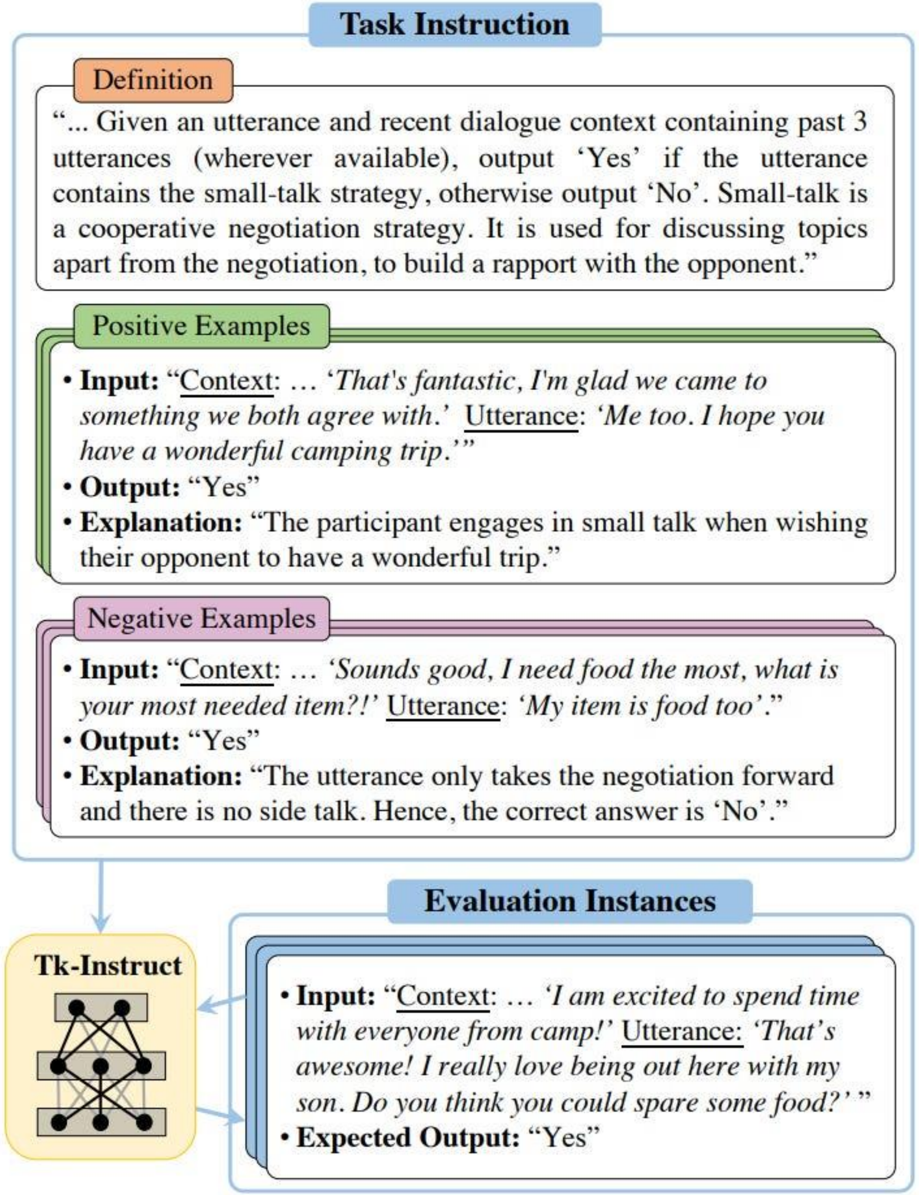


Takeaways

- Language models can be scaled up by increasing model size and training data, and performance follows the power-law relationship.
- Larger models develop emergent abilities that small models do not exhibit, such as in-context learning.
- In-Context Learning (ICL) is mostly impacted by input-output format and text distribution, and less by label accuracy.
- Increasing examples in ICL may improve performance, but the optimal number of examples vary; model performance is sensitive to the order of examples.

Next Class

- Post-training
- Stage 1: Instruction Tuning
  - More training examples
  - More complex tasks
  - Train the model to be flexible to adapt to different kinds of task instructions



Task Instruction

Definition

"... Given an utterance and recent dialogue context containing past 3 utterances (wherever available), output 'Yes' if the utterance contains the small-talk strategy, otherwise output 'No'. Small-talk is a cooperative negotiation strategy. It is used for discussing topics apart from the negotiation, to build a rapport with the opponent."

Positive Examples

- **Input:** "Context: … *'That's fantastic, I'm glad we came to something we both agree with.'* Utterance: *'Me too. I hope you have a wonderful camping trip.'*"
- **Output:** "Yes"
- **Explanation:** "The participant engages in small talk when wishing their opponent to have a wonderful trip."

Negative Examples

- **Input:** "Context: … *'Sounds good, I need food the most, what is your most needed item?!'* Utterance: *'My item is food too'*."
- **Output:** "Yes"
- **Explanation:** "The utterance only takes the negotiation forward and there is no side talk. Hence, the correct answer is 'No'."

Evaluation Instances

- **Input:** "Context: … *'I am excited to spend time with everyone from camp!'* Utterance: *'That's awesome! I really love being out here with my son. Do you think you could spare some food?'*"
- **Expected Output:** "Yes"