UNIVERSITY of VIRGINIA

In-Context Learning (ICL)

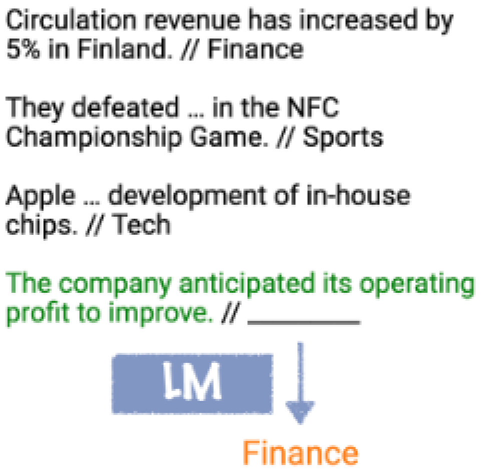
Hamidreza Nabaei and Devaaya Latta

Presentation Outline

- Introduction (What is ICL?)
- Paper 1: Theoretical Foundations ICL as Implicit Bayesian Inference
- Paper 2: What drives ICL performance?
- Paper 3: Exploring Many Shot ICL
- Conclusion
- Discussion

What is In-Context Learning (ICL)?

- In-context learning: Conditioning on examples to make predictions on test examples without optimizing parameters
  - Popularized in original GPT3 paper
  - Works with Large LMs, no optimized of parameters
  - Notable few-shot accuracies on NLP tests (Brown et al. 2020)

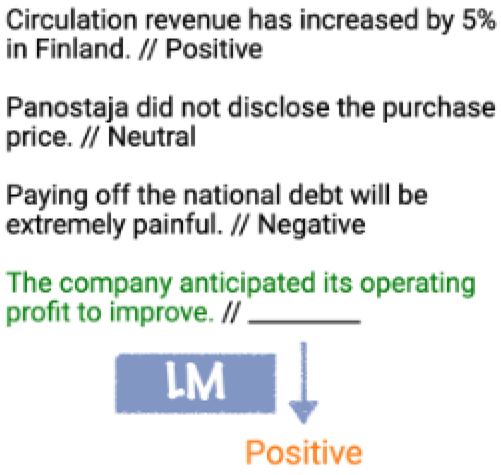


Circulation revenue has increased by 5% in Finland. // Finance

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

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

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



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. // ________

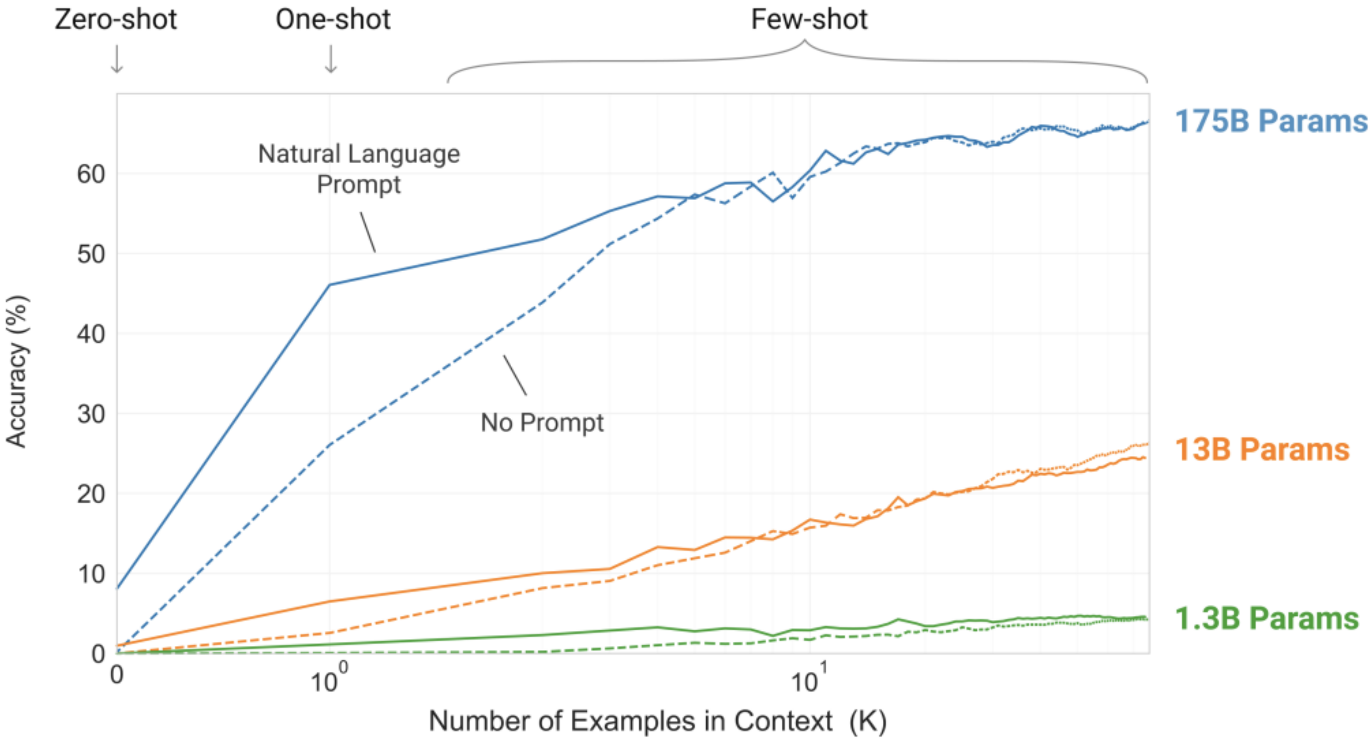
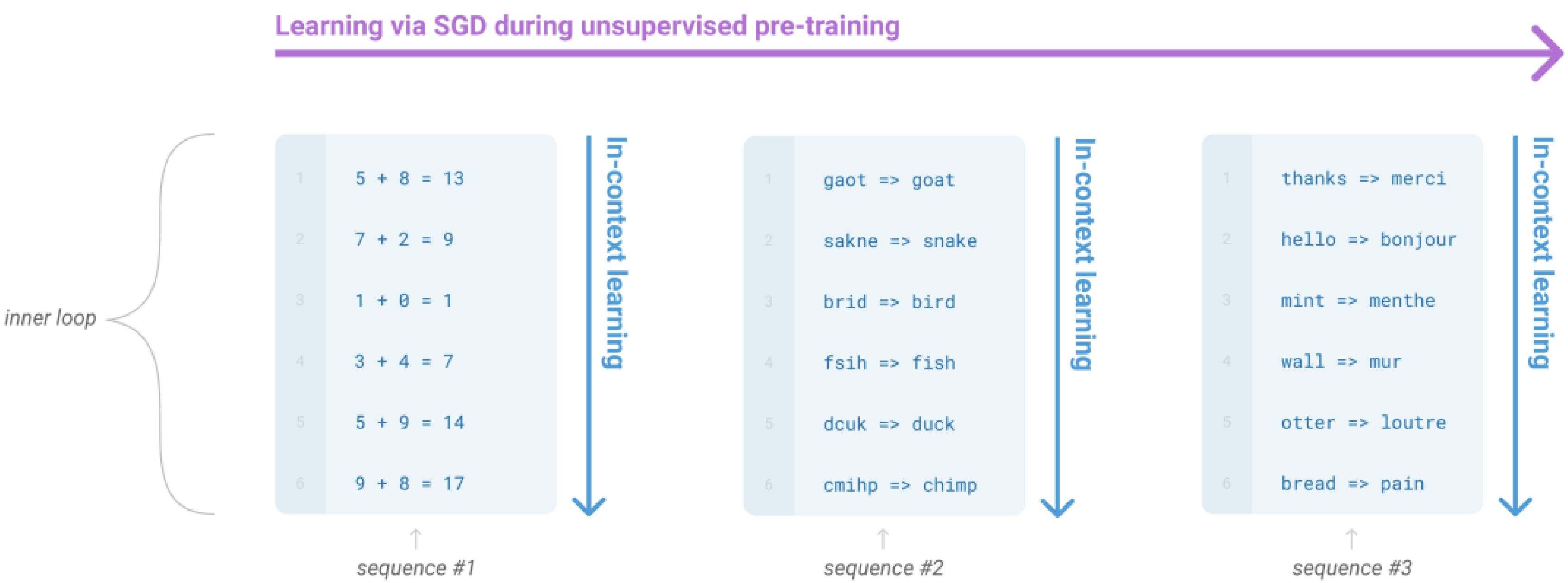


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.



Learning via SGD during unsupervised pre-training

inner loop

```
5 + 8 = 13
7 + 2 = 9
1 + 0 = 1
3 + 4 = 7
5 + 9 = 14
9 + 8 = 17
```

In-context learning

sequence #1

```
gaot => goat
sakne => snake
brid => bird
fsih => fish
dcuk => duck
cmihp => chimp
```

In-context learning

sequence #2

```
thanks => merci
hello => bonjour
mint => menthe
wall => mur
otter => loutre
bread => pain
```

In-context learning

sequence #3

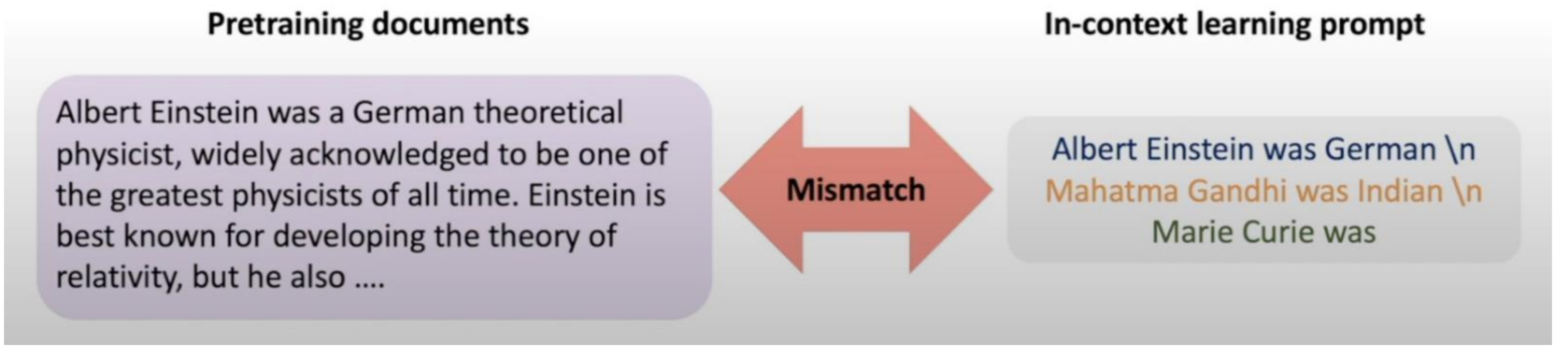
"During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "**in-context learning**" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence" (Brown et al 2020)

What Can ICL do?

- Allows users to quickly build models for a new use cases without fine-tuning and storing new parameters for each novel task.
- On many NLP benchmarks, ICL competitive with models trained with much more labeled data and is state of the art on LAMBADA and TriviaQA
  - ICL is essential for application tasks (app design mockups, website design, spreadsheet novel programming etc.)

Why is ICL Interesting?

- LM not explicitly trained for learning
- Prompts differently formatted than NLP pretraining documents



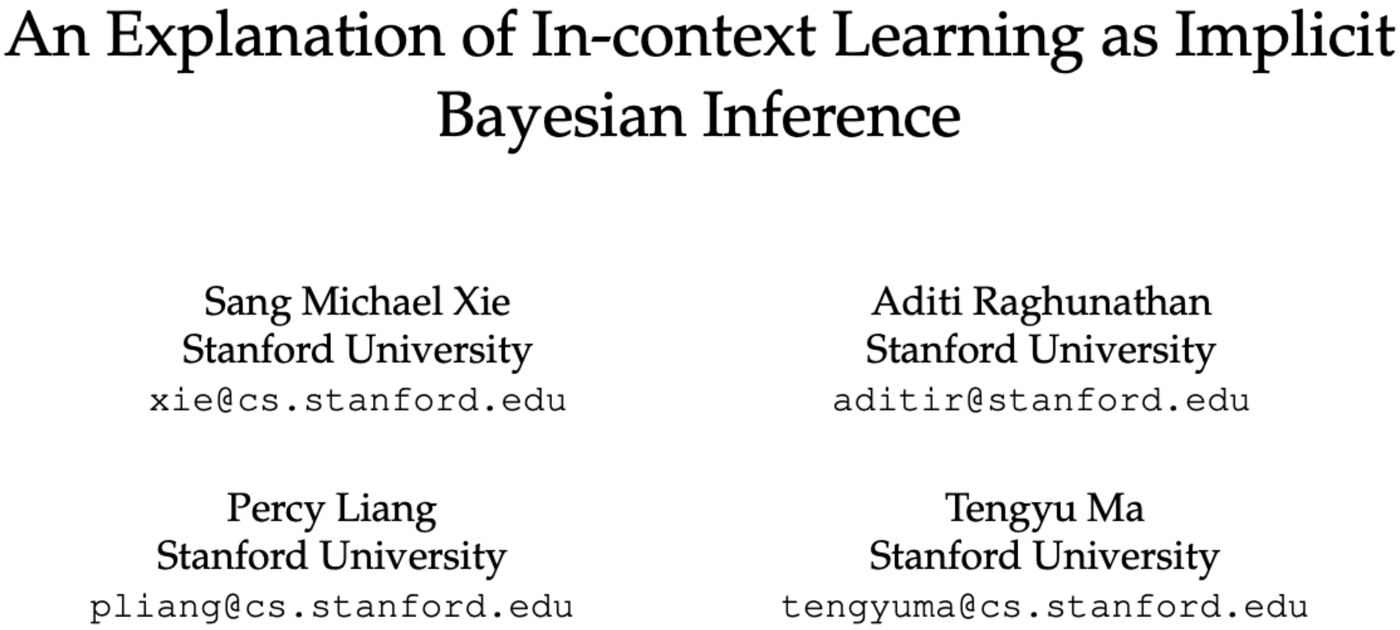
Pretraining documents

Albert Einstein was a German theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is best known for developing the theory of relativity, but he also ....

In-context learning prompt

Albert Einstein was German \n
Mahatma Gandhi was Indian \n
Marie Curie was

Paper 1 (P1): ICL as Implicit Bayesian Inference



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

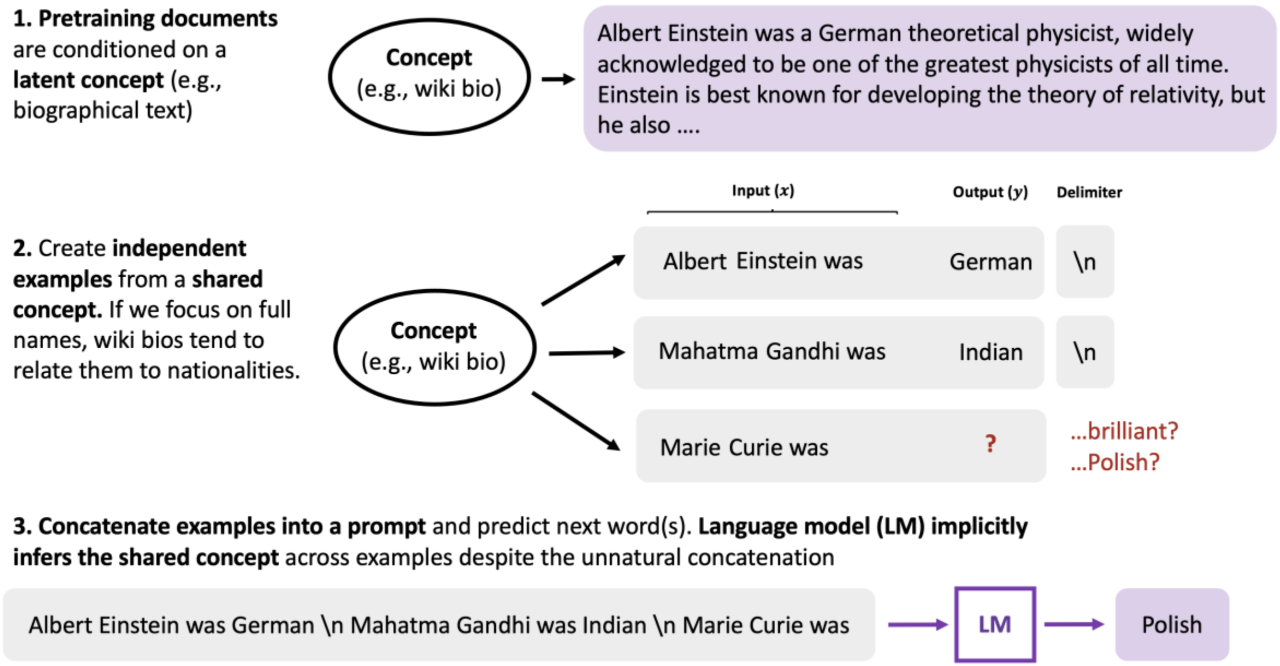
Percy Liang
Stanford University
pliang@cs.stanford.edu

Tengyu Ma
Stanford University
tengyuma@cs.stanford.edu

https://arxiv.org/pdf/2111.02080.pdf

Main Hypothesis: So How Does ICL Work?

- Bayesian Framework (Inference of Latent Concepts)



1. Pretraining documents are conditioned on a **latent concept** (e.g., biographical text)

Concept (e.g., wiki bio)

Albert Einstein was a German theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is best known for developing the theory of relativity, but he also ....

Input (x)	Output (y)	Delimiter
Albert Einstein was	German	\n
Mahatma Gandhi was	Indian	\n
Marie Curie was	?	...brilliant? ...Polish?

2. Create **independent examples** from a **shared concept.** If we focus on full names, wiki bios tend to relate them to nationalities.

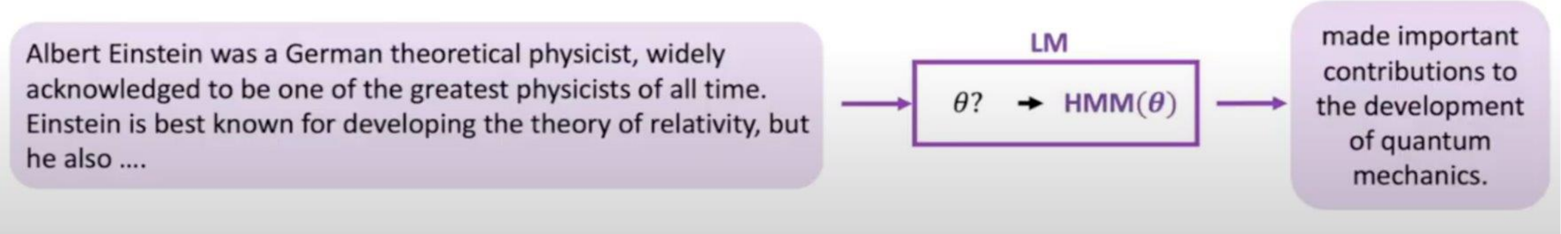
Concept (e.g., wiki bio)

3. Concatenate examples into a prompt and predict next word(s). **Language model (LM) implicitly infers the shared concept** across examples despite the unnatural concatenation

Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was → LM → Polish

P1. Overview (Purpose of Paper)

- Proposes a framework in which LM uses ICL to **locate** a previously learned concept to do ICL task
- Model generates predictions based upon located latent context rather than pre-trained task-specific model
- **Locating** learned capabilities can be mathematically formulized as Bayesian inference

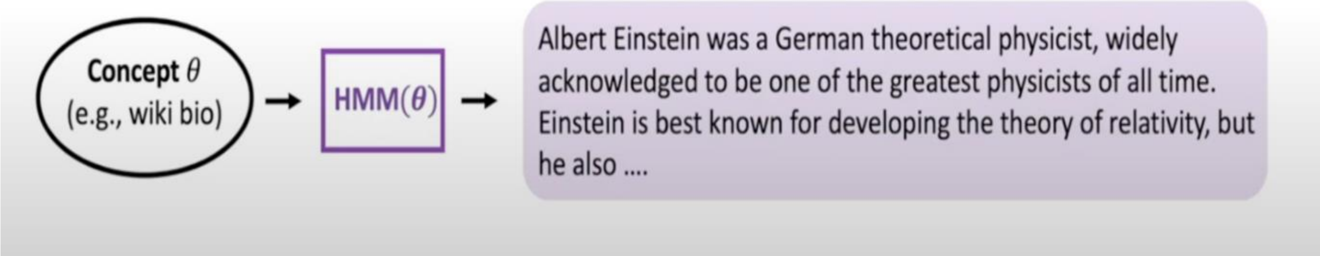


Albert Einstein was a German theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is best known for developing the theory of relativity, but he also ....

made important contributions to the development of quantum mechanics.

P1. Coherence

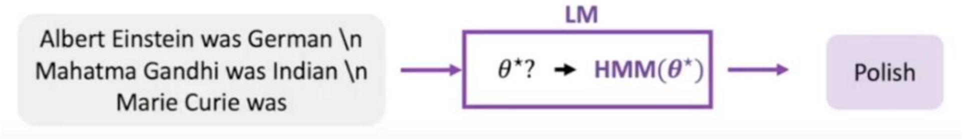
- Since sentences from pretraining documents sharing a concept (long-term coherence)
  - Example Concept θ (wiki bio text)



Albert Einstein was a German theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is best known for developing the theory of relativity, but he also ....

P1. Main Challenge of ICL

- ICL emerges when LM infers shared concept θ from IC examples
- LM works if signal around theta is greater than noise from low probability transitions

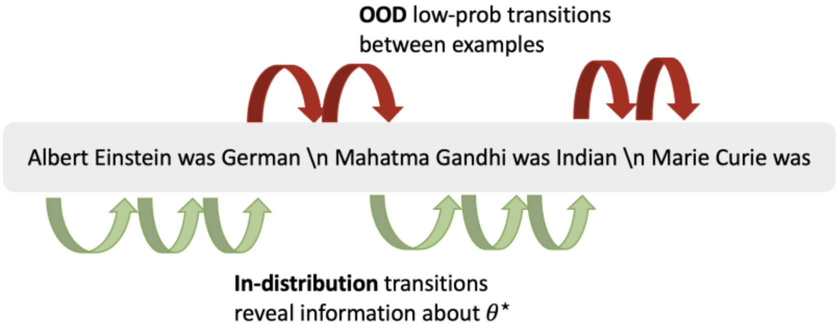


Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was

LM

$\theta^{\star}? \rightarrow \text{HMM}(\theta^{\star})$

Polish



OOD low-prob transitions between examples

Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was

In-distribution transitions reveal information about $\theta^{\star}$

P1. Bayesian Inference Framework

- **Pretraining Distribution (p):** assume that pretraining documents are generated by first sampling of latent concept, and then generated by conditioning on latent concept.
  - Data/LM large enough so that LM fits pretraining distribution exactly

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

P1. Bayesian Inference Framework

- Question of ICL is characterizing p(output|prompt) under the pretraining distribution
- "If p(concept|prompt) concentrates on the prompt concept with more examples, then the LM learns via marginalization by “selecting” the prompt concept."

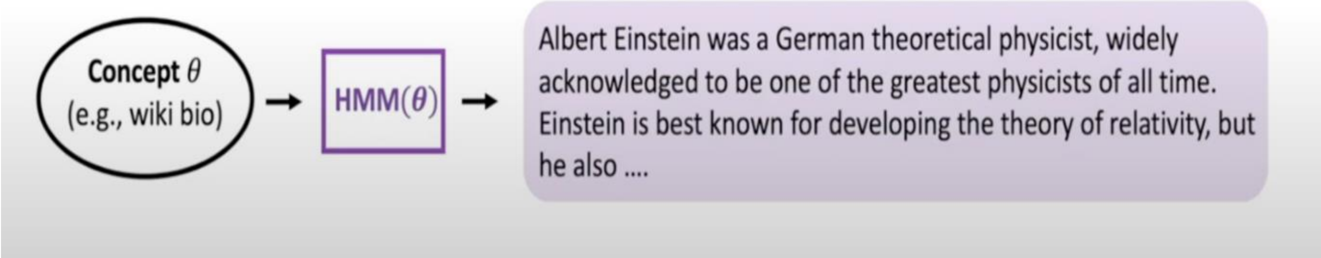
$$p(\text{output}|\text{prompt}) = \int_{\text{concept}} p(\text{output}|\text{concept}, \text{prompt})p(\text{concept}|\text{prompt})d(\text{concept}).$$

P1. Hidden Markov Model (HMM)

- Each pretraining document is a length T sequence sampled by

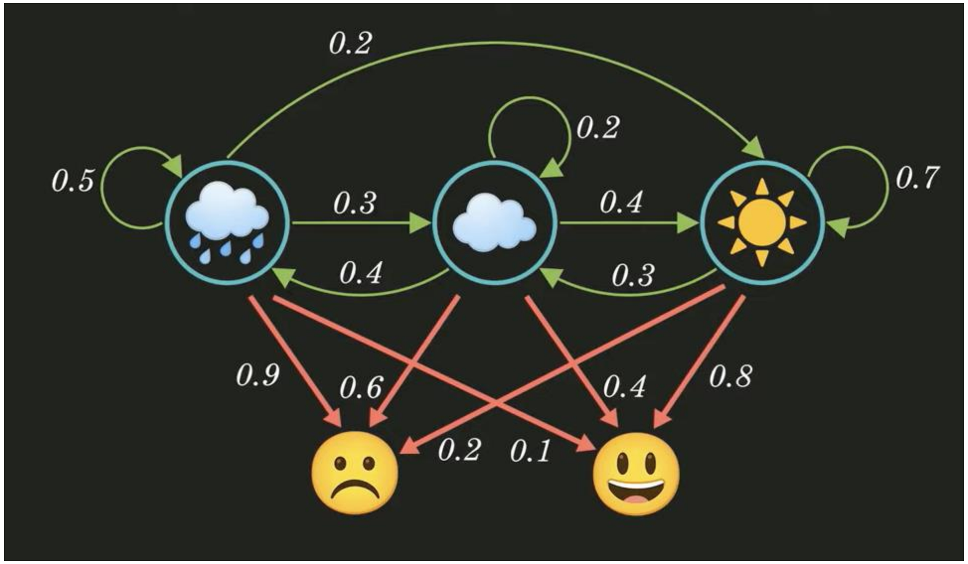
$$p(o_1, \ldots, o_T) = \int_{\theta \in \Theta} p(o_1, \ldots, o_T | \theta) p(\theta) d\theta,$$

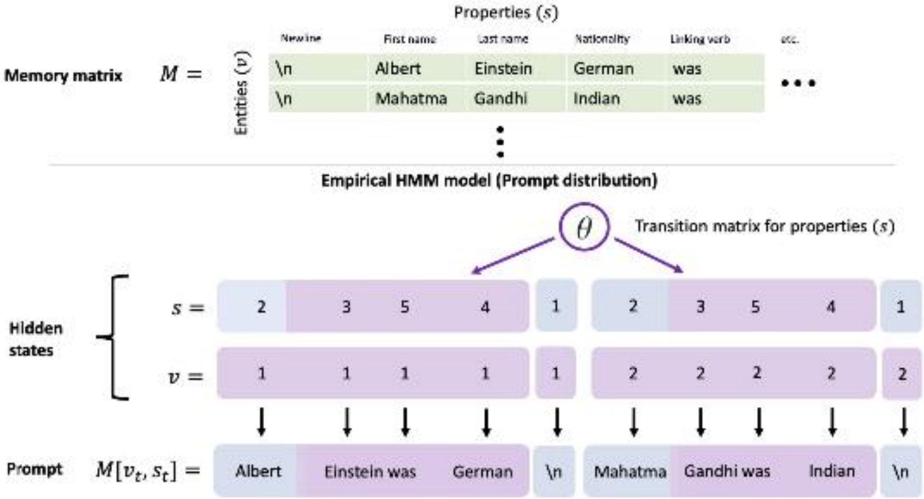
  - Assumption of paper is that p(o1,...oT) is defined by Hiden Markov Model (HMM). The concept Theta determines transition probability matrix of HMM hidden states



Albert Einstein was a German theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is best known for developing the theory of relativity, but he also ....

HMMs




Properties (s)

Memory matrix $M =$ Entities (v)

New line	First name	Last name	Nationality	Linking verb	etc.
\n	Albert	Einstein	German	was	...
\n	Mahatma	Gandhi	Indian	was	

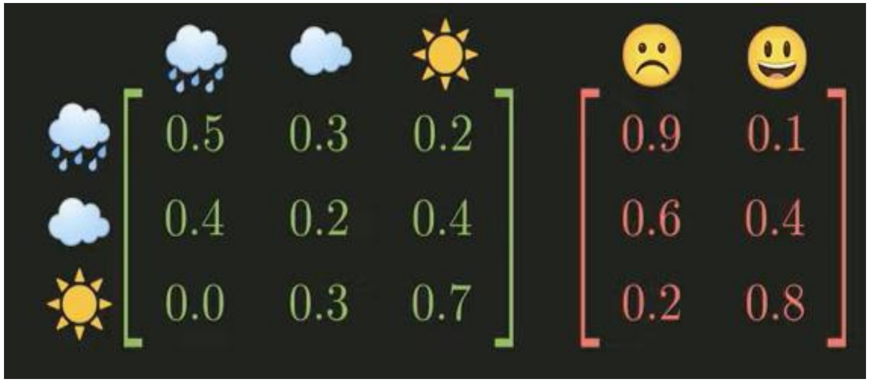
⋮

Empirical HMM model (Prompt distribution)

θ Transition matrix for properties (s)

Hidden states

$s =$	2	3	5	4	1	2	3	5	4	1
$v =$	1	1	1	1	1	2	2	2	2	2
Prompt $M[v_t, s_t] =$	Albert	Einstein was		German	\n	Mahatma	Gandhi was		Indian	\n



	🌧️	☁️	☀️	☹️	😃
🌧️	0.5	0.3	0.2	0.9	0.1
☁️	0.4	0.2	0.4	0.6	0.4
☀️	0.0	0.3	0.7	0.2	0.8

$$\arg\max_{y} \; p(y|S_n, x_{\text{test}}) \rightarrow \arg\max_{y} \; p_{\text{prompt}}(y|x_{\text{test}})$$

- The prompt consists of a sequence of training examples (Sn) followed by the test example xtest: [Sn, xtest] = [x1, y1, odelim, x2, y2, odelim, . . . , xn, yn, odelim, xtest] ∼ pprompt.
- More examples means more signals for Bayesian inference which means smaller error
- As n goes to infinity the incontext prediction asymptotically gets to the expected error

P1. Testing Framework (GINC)

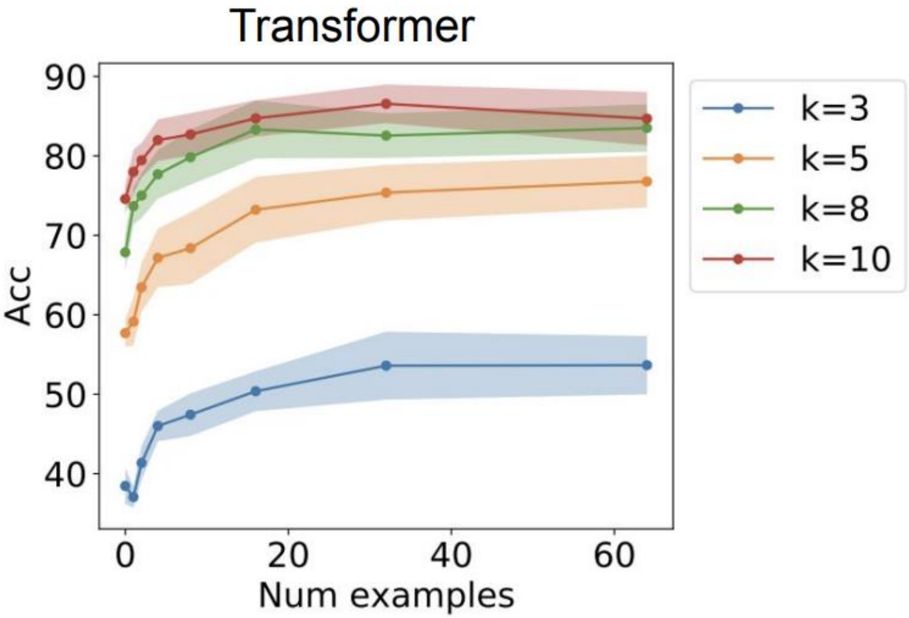
- Synthetic (non-human readable) dataset of sequences generated by admixtures of HMMs
- Designed to mimic different knowledge retrieval tasks
- **Pre-Training: uniform mixture of HMMs over 5 concepts using 1000 pretraining documents**

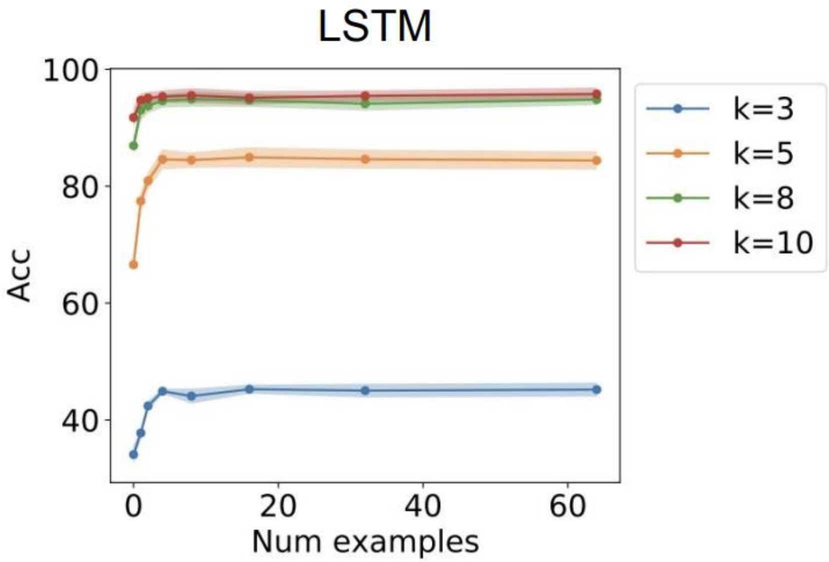
P1. Tests

Accuracy is enhanced with:

- Number of examples and parameters

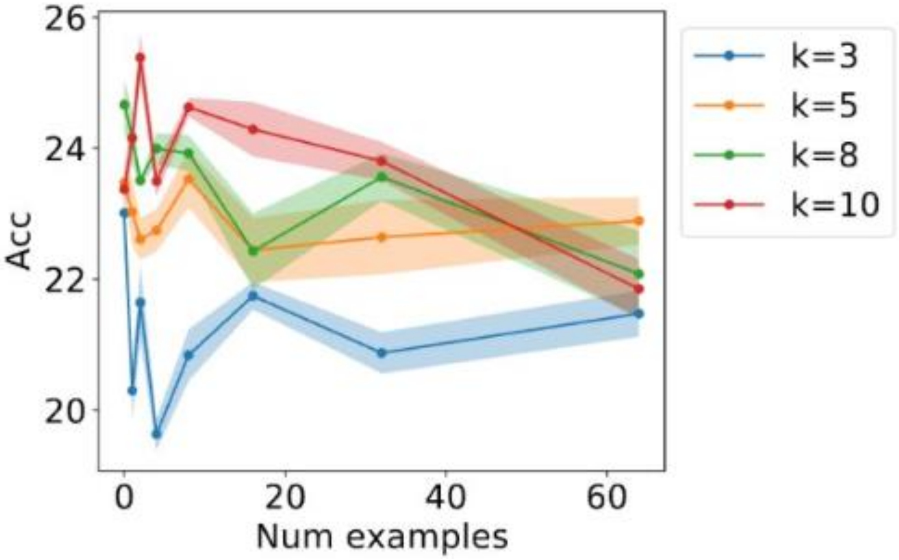
- LSTMs instead of Transformers


Transformer


LSTM

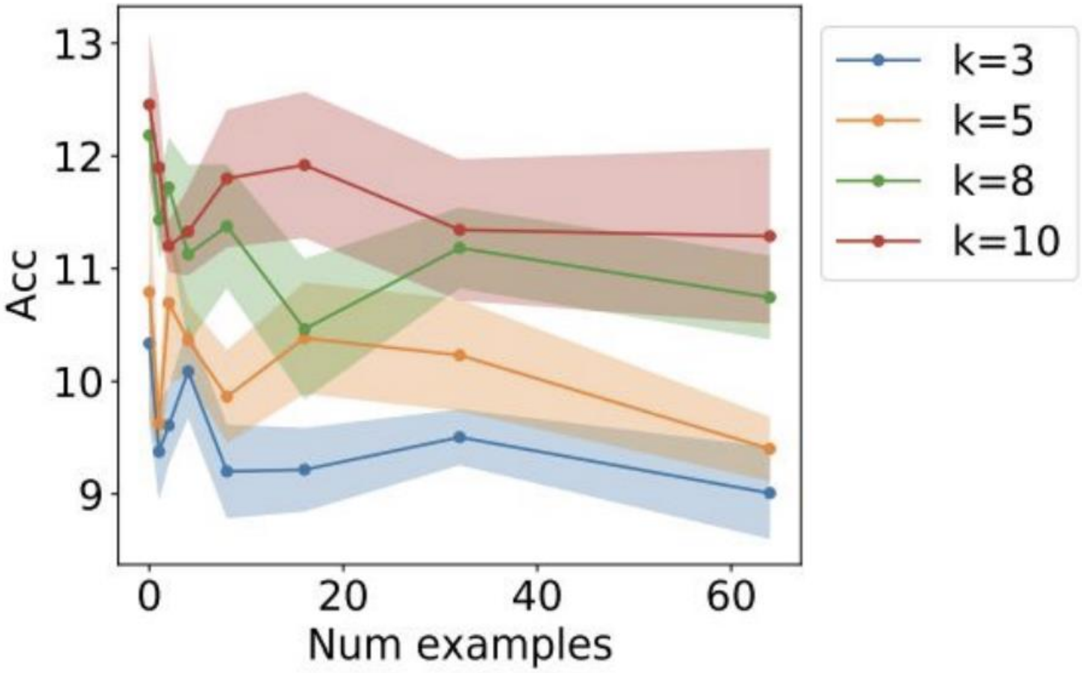
P1. Is HMM assumption important? (Ablation Study)

When pretrained on only one concept, ICL fails



P1. Is HMM assumption important? (Ablation Study)

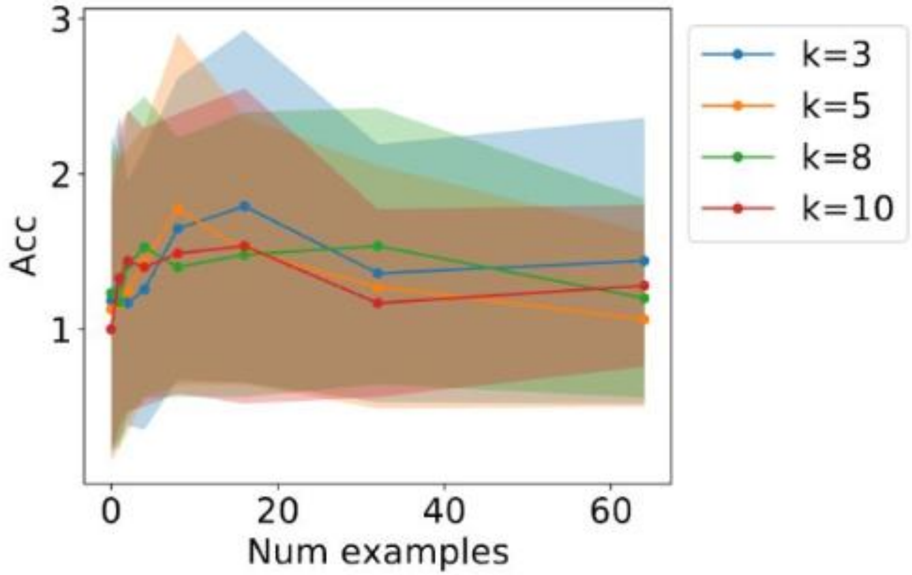
When prompts are from random, not before seen concepts, ICL fails



Pretraining of 4 layer Transformer on only one concept

P1. Is HMM assumption important? (Ablation Study)

When pretraining data is made from random transitions, ICL fails



Summary

- In context learning emerges when texts are modeled as HMMs with sufficiently distinguishable concepts
- Language models are recognizing previously seen concepts instead of learning patterns live

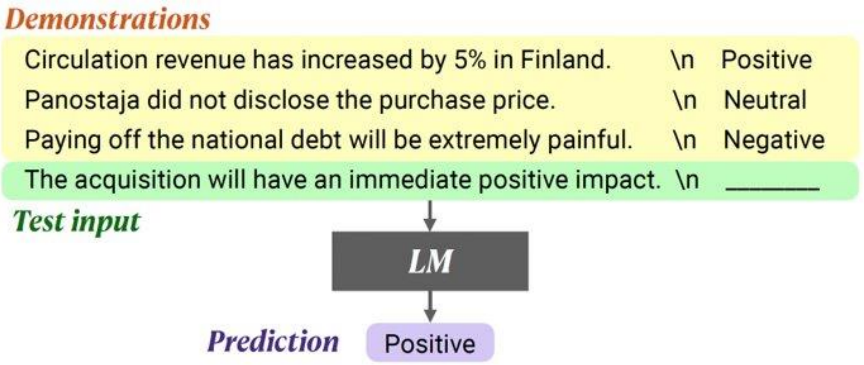
Assumption Check

- Paper does not really study interactions with true natural language.
- The HMM model can't really cover generalizations to truly novel tasks

Paper 2: "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?"

In Context Learning

- A new paradigm in NLP where LLMs **make predictions based on context** augmented with just few training examples(demonstrations) --> lower computation costs
- Then **LLMs extract patterns from the examples** provided in the context and use them to complete complex NLP tasks by conditioning on examples (demonstrations) without fine-tuning.
- Also known as **few-shot learning**, is where a few examples of input-label pairs are supplied to the model as part of the prompt.
- *A frozen LM performs a task only by conditioning on the prompt text.*
- **Definition of ICL:** Learning from input-output examples without weight updates.



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

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

Figure 2: An overview of in-context learning. The demonstrations consist of k input-label pairs from the training data ($k = 3$ in the figure).

Terminology

Prompt	The input text that specifies the task.
Demonstration	Example(s) included in the prompt to show the model how to perform the task.
Shot	The number of demonstrations/examples given in the prompt (zero-shot, one-shot, few-shot, many-shot).

Without Demonstration (0-shot):

- *Prompt:* “Translate ‘Hello’ to French.”
- *Model Response:* “Bonjour.”

With Demonstration (1-shot):

- *Prompt: “Translate the following to French: ‘I am happy.’*
  Example: ‘Good morning’ → ‘Bonjour’. Now translate ‘I am happy’.”
- *Model Response: “Je suis heureux.”*

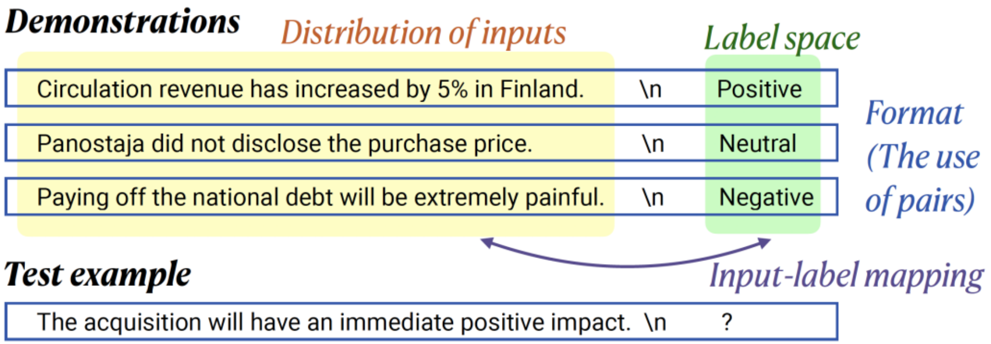
Demonstration

1. Input-label mapping: The process of associating each input with the correct label or output in a dataset.

2. Distribution of input: how the text is structured, organized, or presented

3. Label space: The set of all possible labels/outputs/categories that can be assigned to an input.

4. Format: The structure or representation of the input data and its corresponding label.



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 ?

Figure 7: Four different aspects in the demonstrations: the input-label mapping, the distribution of the input text, the label space, and the use of input-label pairing as the format of the demonstrations.

Experiment setup

“To test their hypothesis, the researchers:

- 12 LMs (GPT-3, GPT-J, etc.)
- 2 **inference** methods: **direct and channel**
- **Evaluation** : 26 datasets that are true-low resource datasets and include GLUE and SuperGLUE and covers diverse domains of science, social media, finance and …
- 16 **demonstrations**
- Run 5-times
- **Classification** and **Multi-choice** tasks
- **Compared ICL with gold labels vs. random labels vs nothing**

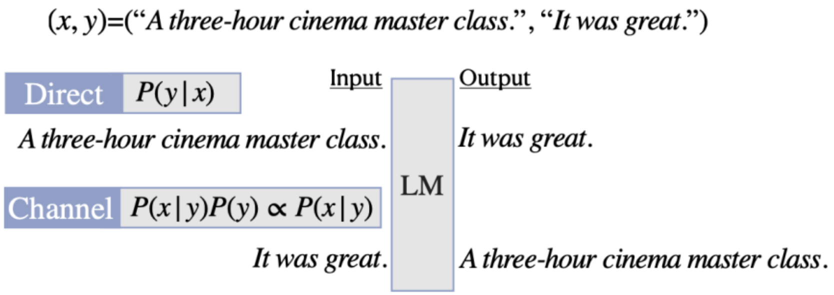

(x, y)=(“*A three-hour cinema master class.*”, “*It was great.*”)
Direct $P(y|x)$
Input
Output
A three-hour cinema master class.
It was great.
LM
Channel $P(x|y)P(y) \propto P(x|y)$
It was great.
A three-hour cinema master class.

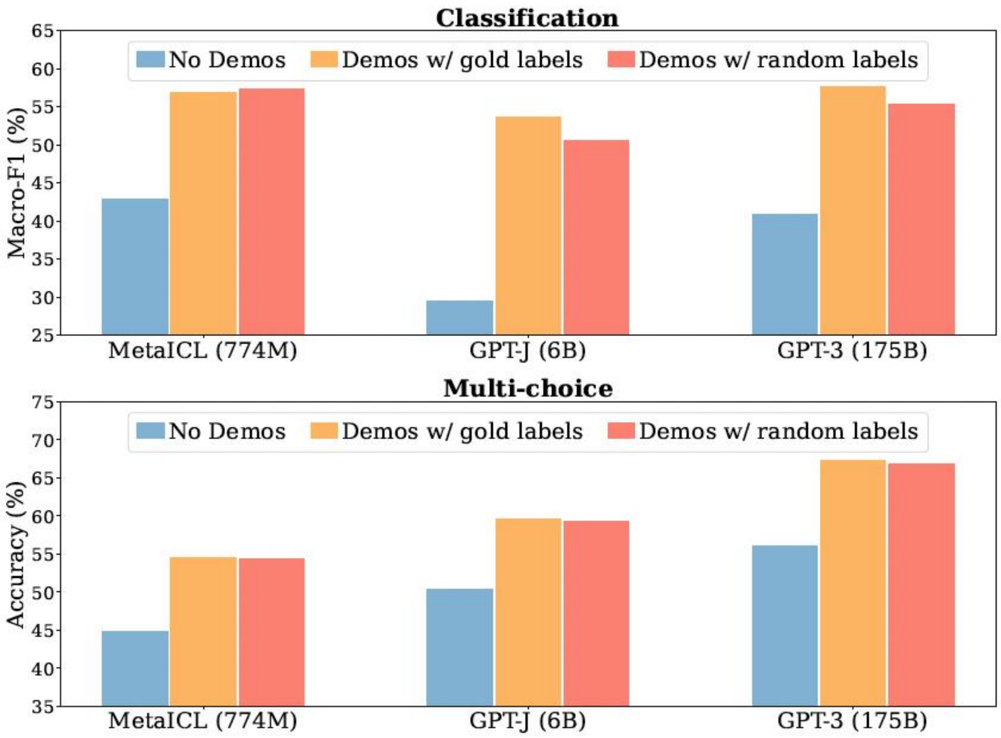
Figure 1: An illustration of the direct model and the channel model for language model prompting in the sentiment analysis task.

Role of Input-Label Mapping

- In-context learning works even with random labels. (**don’t rely on the correct input-label mapping.**)
- **Randomly assigned labels does not decrease model’s performance too much**

Gold labels
X1 y1
X2 y2
X3 y3

Random labels
X1 y2
X2 y3
X3 y1



GOLD vs Random vs NON

1. **Gold labels (correct answers)**
2. **Random labels (incorrect answers assigned randomly, the same label space)**
3. **No demonstrations (zero-shot learning)**

- Replacing gold labels with random labels had minimal impact on model performance

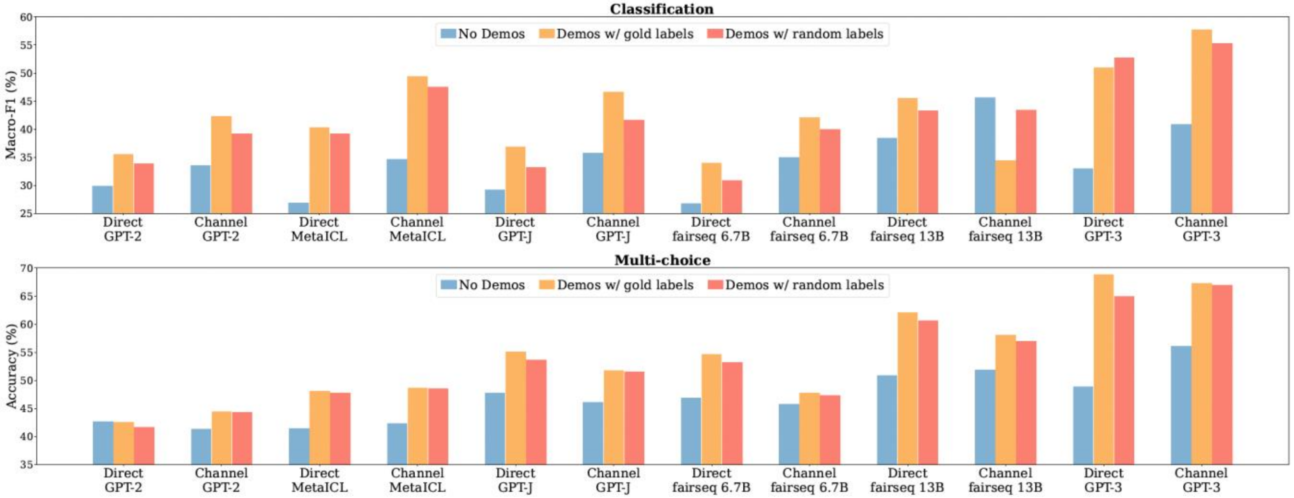


Figure 3: Results when using no-demonstrations, demonstrations with gold labels, and demonstrations with random labels in classification (top) and multi-choice tasks (bottom). The first eight models are evaluated on 16 classification and 10 multi-choice datasets, and the last four models are evaluated on 3 classification and 3 multi-choice datasets. See Figure 11 for numbers comparable across all models. **Model performance with random labels is very close to performance with gold labels** (more discussion in Section 4.1).

GOLD percentages effects

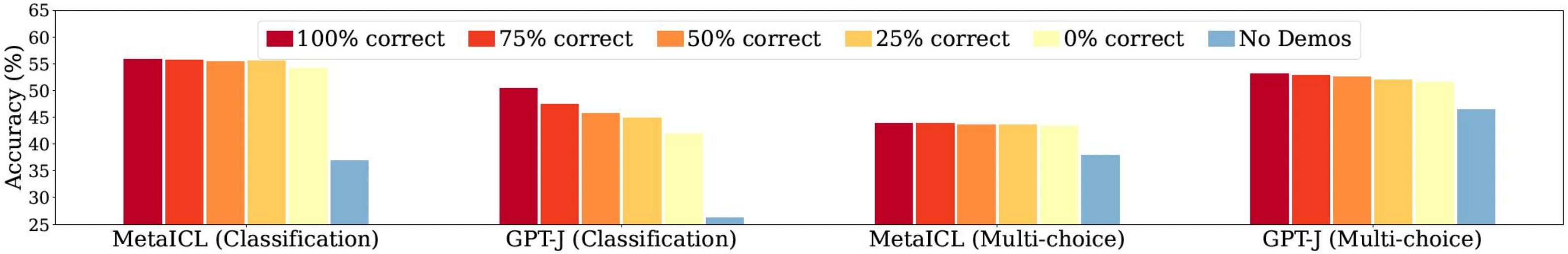


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

- **0% correct labels performance is not too much different with the 100% correct**

GOLD percentages effects

- Increasing the number of demonstrations increase the performance, not too much difference between random and gold

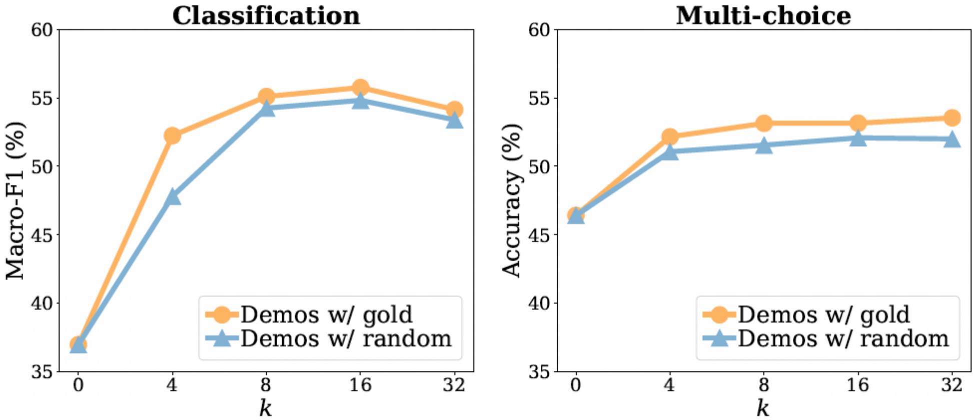


Figure 5: Ablations on varying numbers of examples in the demonstrations (k). Models that are the best under 13B in each task category (Channel MetaICL and Direct GPT-J, respectively) are used.

Different templates

- **Minimal vs Manual:** It worth nothing that using manual templates does not always outperform using minimal **templates**
- **Increasing** the words in prompt —-> not too much **difference**

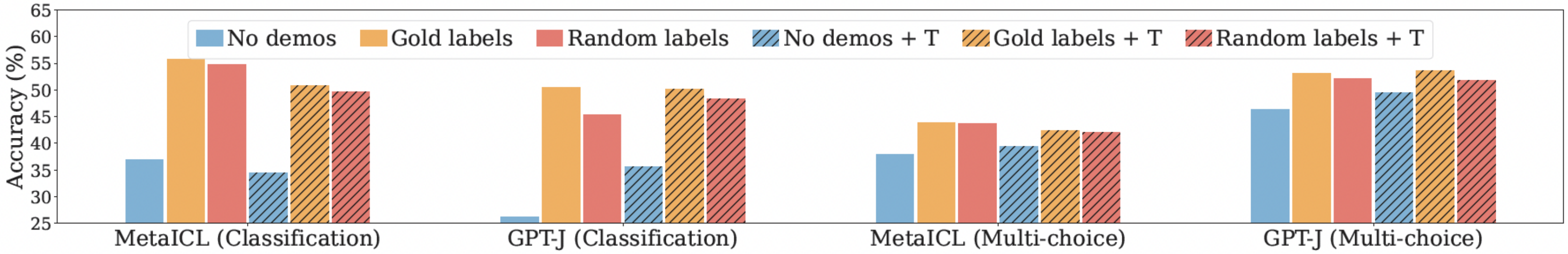


Figure 6: Results with minimal templates and manual templates. '+T' indicates that manual templates are used. Channel and Direct used for classification and multi-choice, respectively.

Role of Distribution

using out-of-distribution inputs instead of the inputs from the training data significantly drops the performance

how the text is structured, organized, or presented

Gold labels	Random labels	OOD
X1 y1	X1 y2	X’1 y2
X2 y2	X2 y3	X’2 y3
X3 y3	X3 y1	X’3 y1

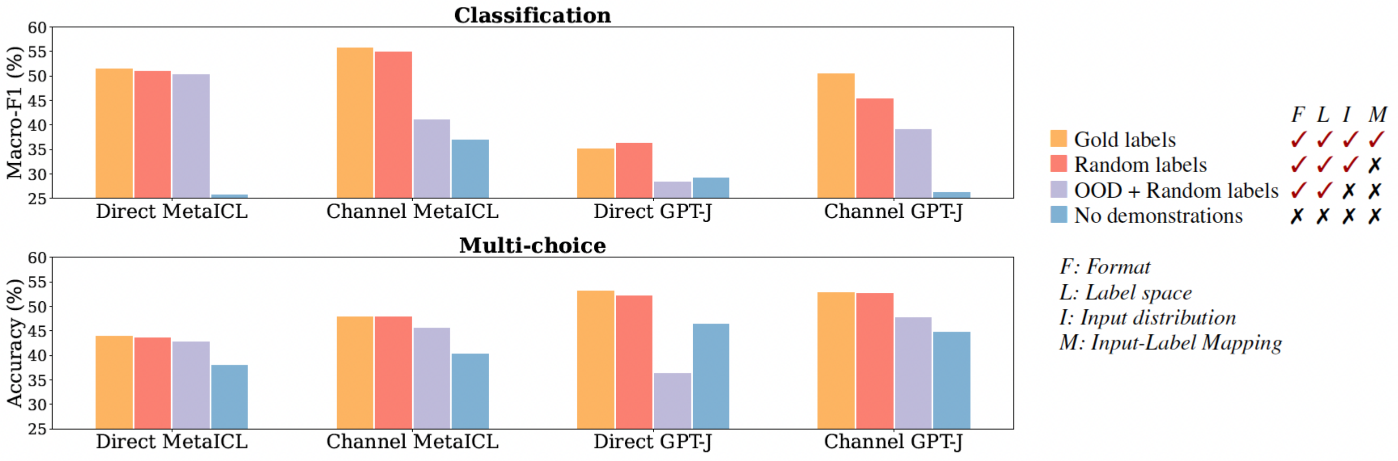


Figure 8: Impact of the distribution of the inputs. Evaluated in classification (top) and multi-choice (bottom). The impact of the distribution of the input text can be measured by comparing 🟥 and 🟪. The gap is substantial, with an exception in Direct MetaICL (discussion in Section 5.1).

Role of Label Space

The model benefits from knowing the types of answers

The set of all possible labels

Gold labels
X1 y1
X2 y2
X3 y3

Random labels
X1 y2
X2 y3
X3 y1

Random English words
X1 y4
X2 y6
X3 y5

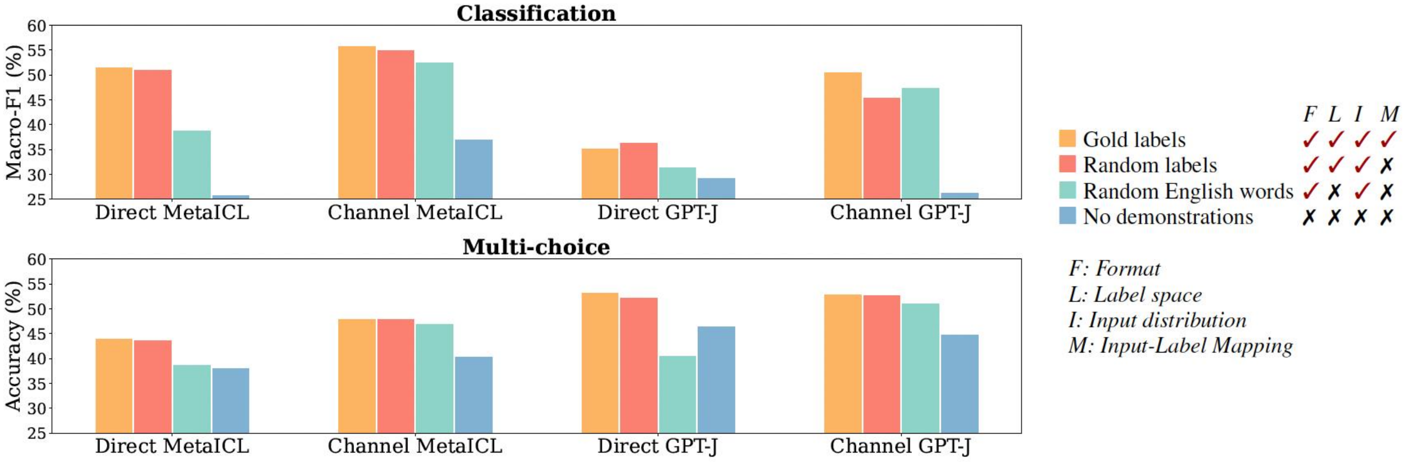


Figure 9: Impact of the label space. Evaluated in classification (top) and multi-choice (bottom). The impact of the label space can be measured by comparing 🟥 and 🟩. The gap is significant in the direct models but not in the channel models (discussion in Section 5.2).

Role of Formatting

keeping the format of the input-label pairs is key.

Format – The way the input and output are structured

Demonstrations with no label

Demonstrations with labels only

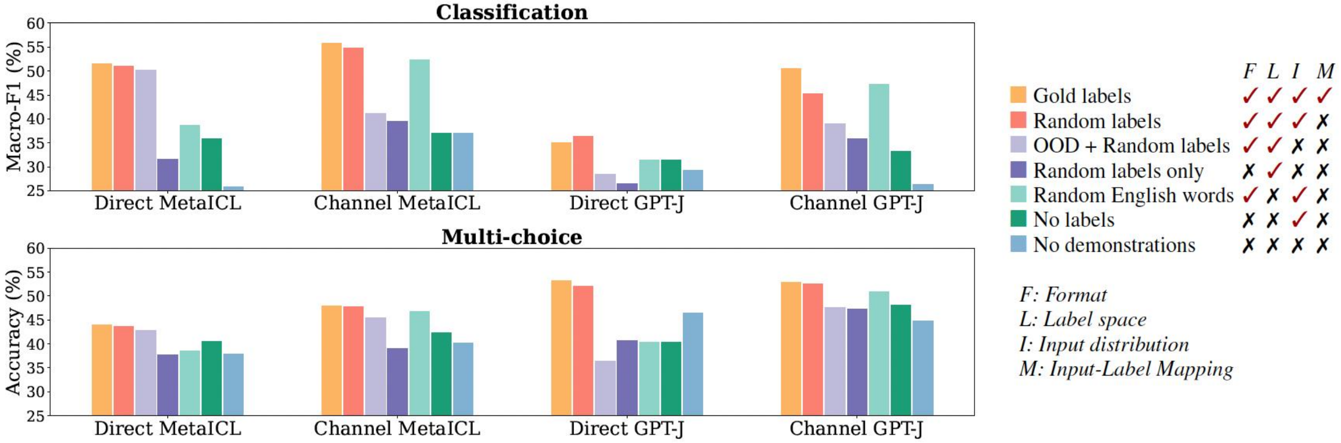


Figure 10: Impact of the format, i.e., the use of the input-label pairs. Evaluated in classification (top) and multi-choice (bottom). Variants of demonstrations without keeping the format (🟪 and 🟩) are overall not better than no demonstrations (🟦). Keeping the format is especially significant when it is possible to achieve substantial gains with the label space but without the inputs (🟪 vs. 🟪 in Direct MetaICL), or with the input distribution but without the labels (🟩 vs. 🟩 in Channel MetaICL and Channel GPT-J). More discussion in Section 5.3.

Role Meta-Training

meta-training encourages the model to exclusively exploit simpler aspects of the demonstrations and to ignore others

MetaICL = is trained with an in-context learning objective

focusing on learning how to use these demonstrations effectively without gradient updates

while training they had demonstrations as input and labels as output so It was trained in this context learning setup

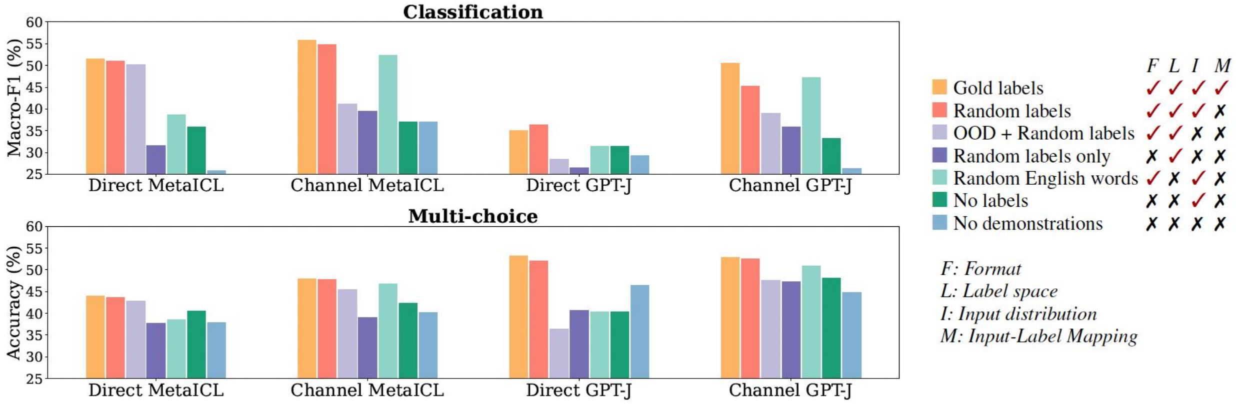


Figure 10: Impact of the format, i.e., the use of the input-label pairs. Evaluated in classification (top) and multi-choice (bottom). Variants of demonstrations without keeping the format (🟪 and 🟩) are overall not better than no demonstrations (🟦). Keeping the format is especially significant when it is possible to achieve substantial gains with the label space but without the inputs (🟪 vs. 🟪 in Direct MetaICL), or with the input distribution but without the labels (🟩 vs. 🟩 in Channel MetaICL and Channel GPT-J). More discussion in Section 5.3.

Key Findings

“**So, if the models don’t rely on the correct input-label mapping.**
what actually helps the model perform better?

1️⃣ **Label space** – The model benefits from knowing the types of answers (e.g., Positive/Negative/Neutral).
2️⃣ **Input distribution** – Seeing real-world examples makes the model better at generalizing.
3️⃣ **Format** – The way the input and output are structured (e.g., question-answer pairs) helps the model recognize patterns.”

So :

1. Ground truth demonstrations are not required (*LMs do not need input-label mapping in demonstrations, instead, it uses the specification of the input & label distribution separately*)
2. Understanding task distribution and label space is key.
3. Future work should focus on improving model adaptability

Implications & Future Work

1. Do LLMs actually ‘learn’ at test time ❓

a model doesn’t necessarily ‘learn’ in the traditional sense—it **doesn’t need correct examples to perform well**. Instead, it just needs to see the right format and input structure(recognizing familiar patterns?)” models are NOT learning specific input-label mappings. Instead, they benefit more from just seeing **structured examples** of inputs and outputs.”

2. Risks(limitations) of using ICL for unseen tasks.

⚠️ if models aren’t really ‘understanding’ tasks, they might fail in unexpected ways.

May not generalize across all tasks and datasets, as some tasks are more sensitive to the use of ground truth labels than others.

3. How does this impact instruction-following models?

A need to explore **how models can improve their learning process** beyond just mimicking patterns.”

4. Why In-Context Learning Works

The previous two observations hence suggest that the performance gains from in-context learning vs zero-shot learning is due to the specification of the input space and label space to the model, and not because the model actually tries to learn from the supplied input-label pairs.

In fact, based on the results the model largely ignores the correspondence of the input-label pairs, and instead uses its own priors during pretraining for the output.

Many-Shot In-Context Learning

a method in NLP where a model learns(is trained to use) from a large number of examples (or “shots”) of a task context

how scaling the number of shots affects ICL performance on a wide variety of tasks ?

1. Many-Shot ICL

- Explores hundreds or thousands of examples for

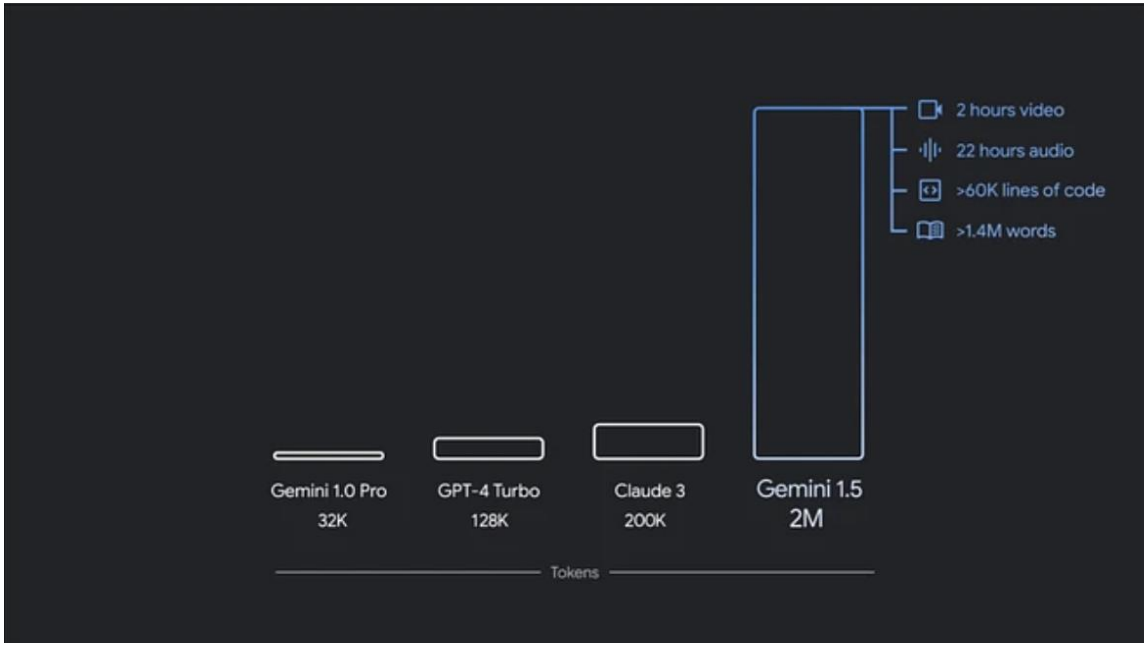
  - Better task specification (Significant performance gains across tasks)

  - Reducing reliance on fine-tuning.

  - Overcome limited context windows (GPT-3 to Gemini 1.5 Pro with 1M tokens).

Context length total amount of text (in tokens) that the model can process at once.

- **Challenges**: Limited by the **need for large numbers of high-quality human-generated examples.**



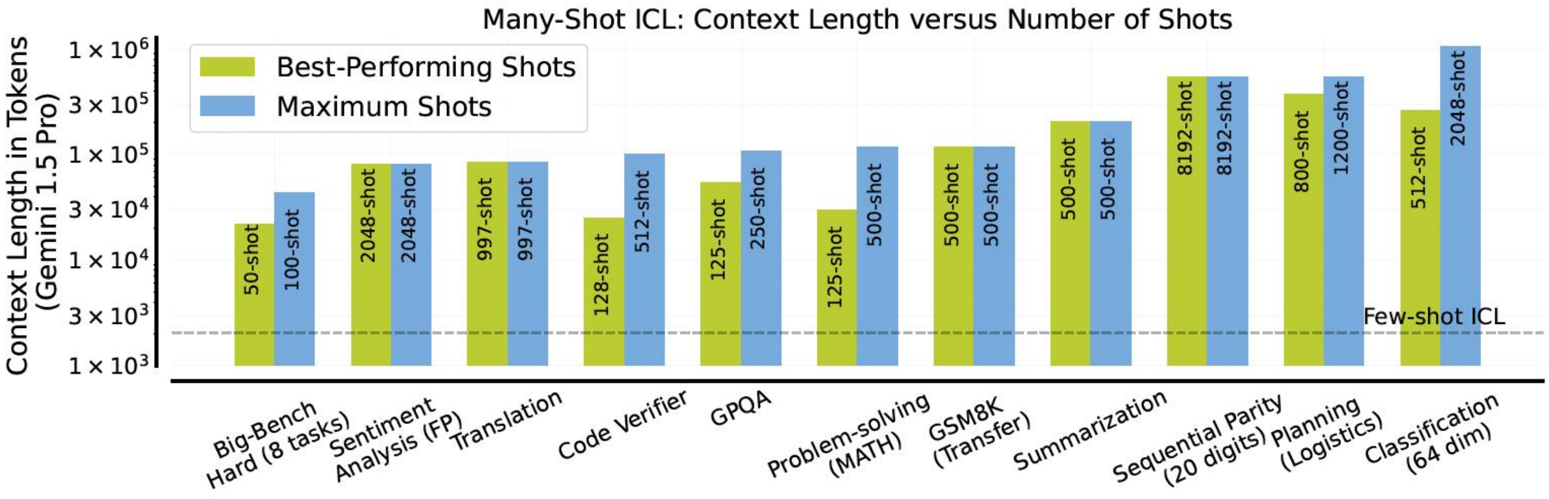
Context length of leading foundation models compare with Gimni1.5's 2 million token

How Many Shots is "many-shot"?

- It is different For different tasks, for some like hundreds and some like thousands

- The graph compares **best-performing shots** (those where the model did the best) with **maximum shots** (the largest number of examples tested).

- For most tasks, the model performed best when given a **larger number of examples (shots)**, but giving too many examples can sometimes reduce performance, especially when the model’s context length limit is exceeded.



Does Many-Shot ICL Improve performance? Yes

Gemini 1.5 Pro: Evaluated across multiple tasks with hundreds to thousands of shots.

- **Key takeaway:** Many-shot ICL significantly outperforms few-shot ICL on complex reasoning tasks.
- **Optimal** number of shots to achieve maximum performance is typically between 100,000 to 1 million tokens -->

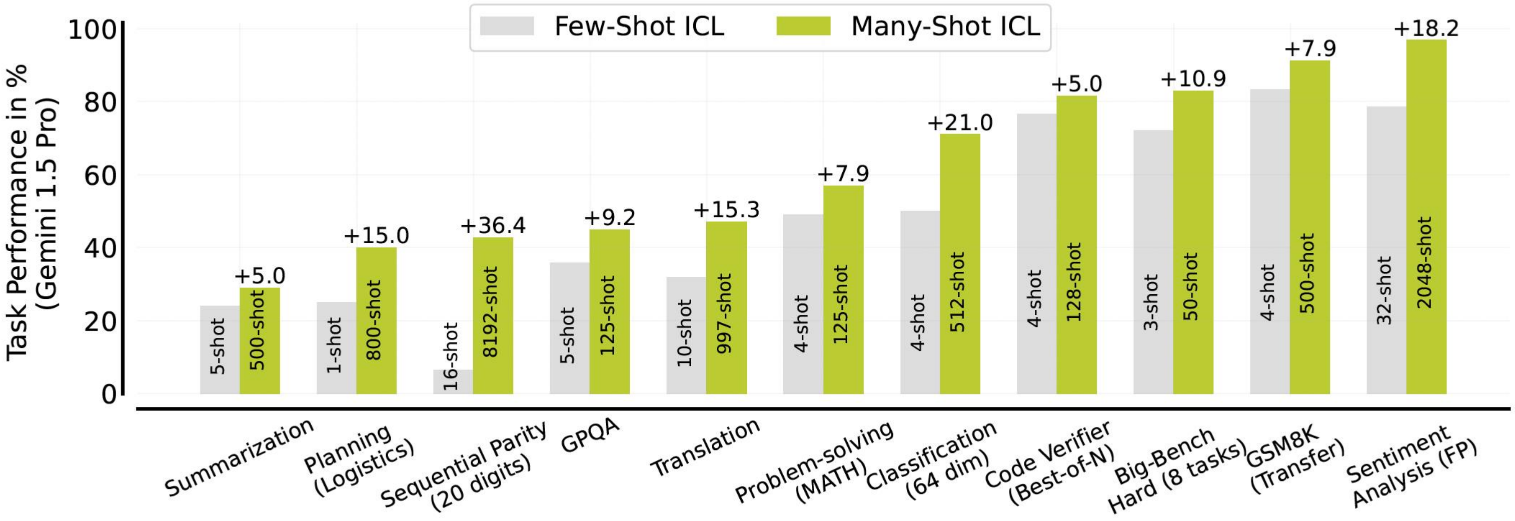
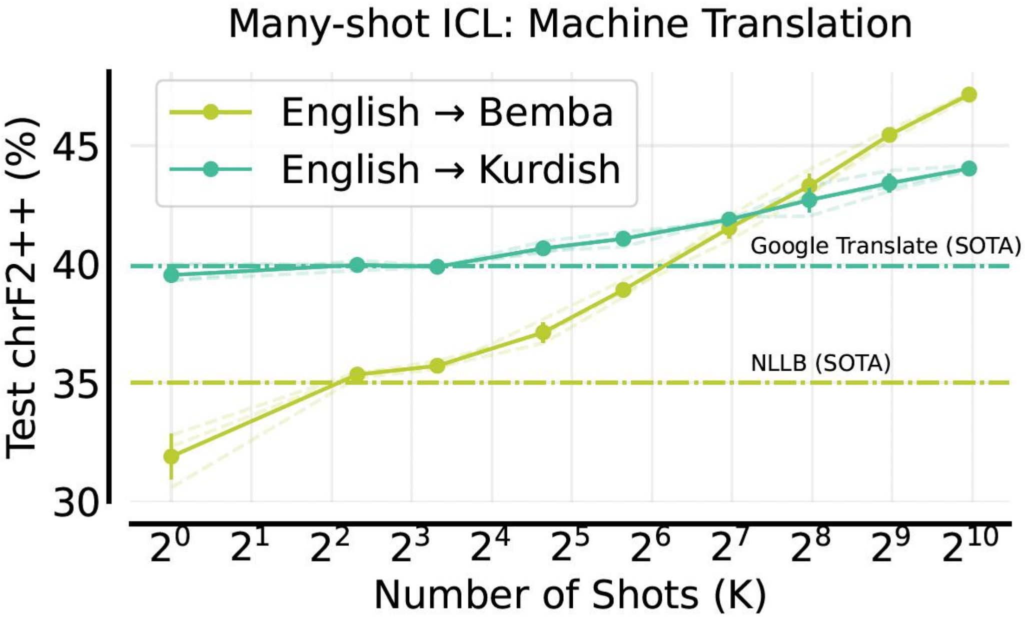


Figure 1 | **Many-shot vs Few-Shot In-Context Learning** (ICL) across several tasks. Many-shot ICL consistently outperforms few-shot ICL, particularly on difficult non-natural language tasks. Optimal number of shots for many-shot ICL are shown

Machine Learning on low-resource Languages

Particular task : Translation from English to low-resource languages (Bemba, Kurdish).

- **Findings**: Many-shot ICL outperforms SOTA and Google Translate with up to 1,000 shots.
- **Performance Gains**: 15.3% improvement on Bemba and 4.5% on Kurdish.

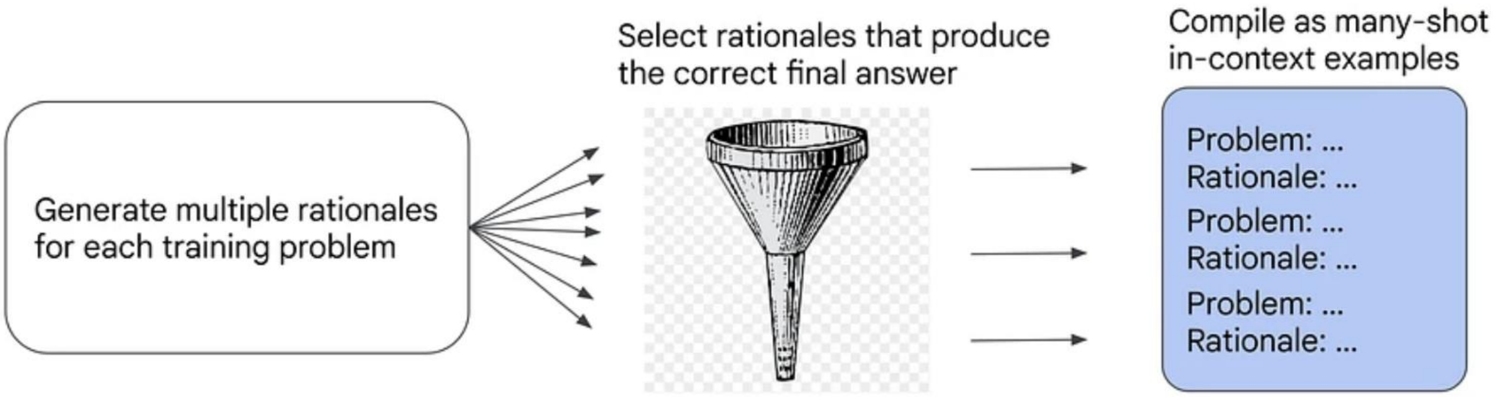


2. Many-Shot Learning without human written Rationales

Replaces human-written rationales with Model-generated ones for problem-solving

- How to run the Many-shot ICL without human rationals?
- Human-written rationales or demonstrations can be expensive to collect, can we do without?

- **Approach: Using Reinforces ICL we can generate model-generated rationales**


Generate multiple rationales for each training problem
Select rationales that produce the correct final answer
Compile as many-shot in-context examples
Problem: ...
Rationale: ...
Problem: ...
Rationale: ...
Problem: ...
Rationale: ...

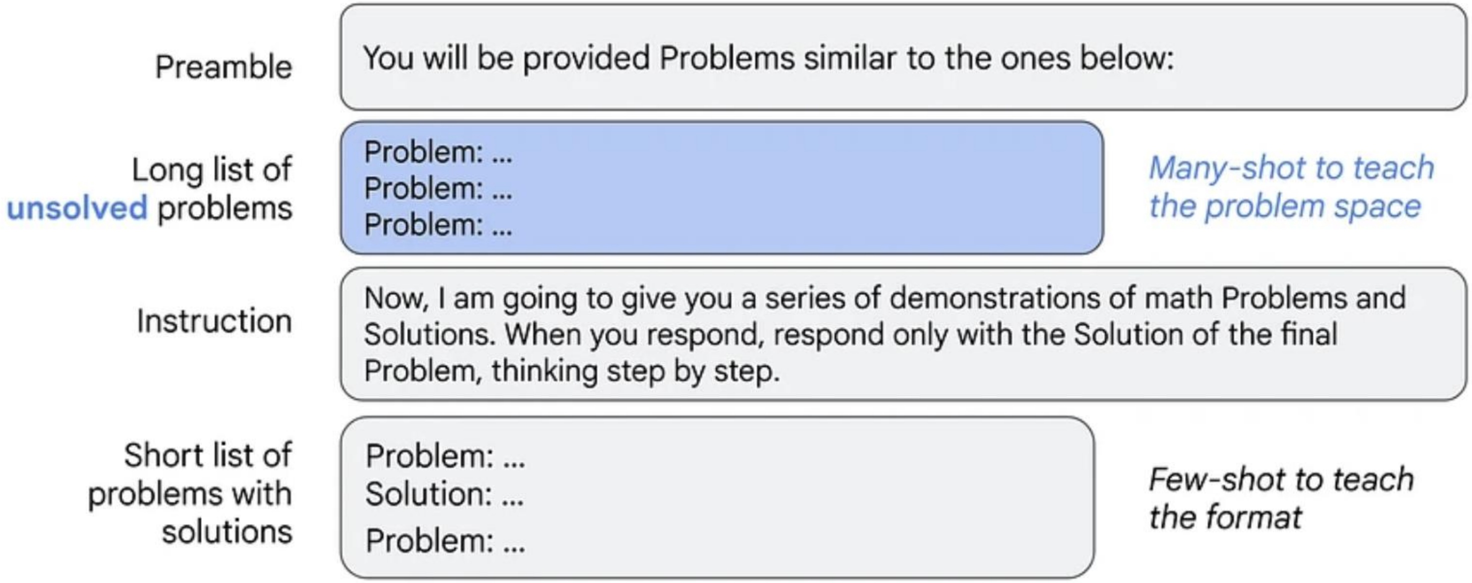
Using model-generated rationales or only problems can reduce the dependence of many-shot ICL on human-generated data.

Another Approach:

Unsupervised ICL: Eliminates rationales, Using only domain-specific inputs.

Get rid of Rationales Entirely

- What if we just put the problem in the context window?
- Providing some problems, then asking for the solution of the problems



Preamble

You will be provided Problems similar to the ones below:

Long list of unsolved problems

Problem: ...
Problem: ...
Problem: ...

Many-shot to teach the problem space

Instruction

Now, I am going to give you a series of demonstrations of math Problems and Solutions. When you respond, respond only with the Solution of the final Problem, thinking step by step.

Short list of problems with solutions

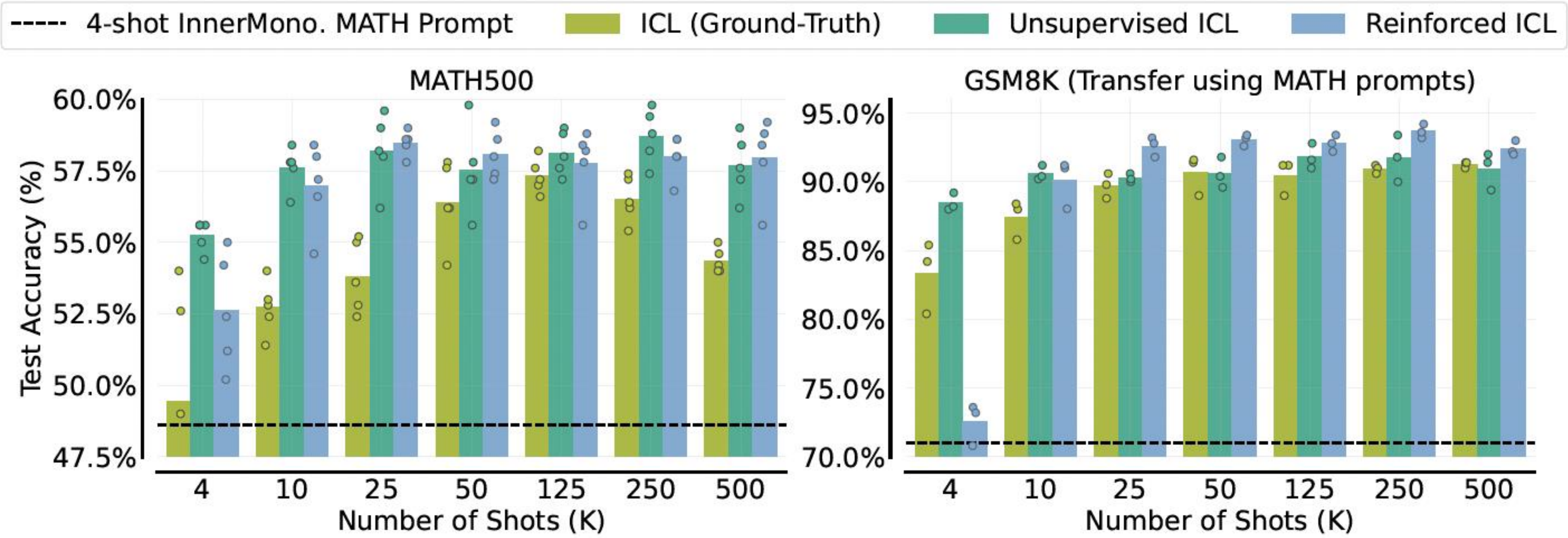
Problem: ...
Solution: ...
Problem: ...

Few-shot to teach the format

Reinforced and Unsupervised ICL Evaluation for Reasoning tasks

- **Results:** Both methods are effective for complex reasoning tasks, with Reinforced ICL being broadly more effective and outperforms ICL with human-written solutions!

- Math and GSM8k : two datasets commonly used in ML and NLP for evaluating models on mathematical reasoning and problem-solving.



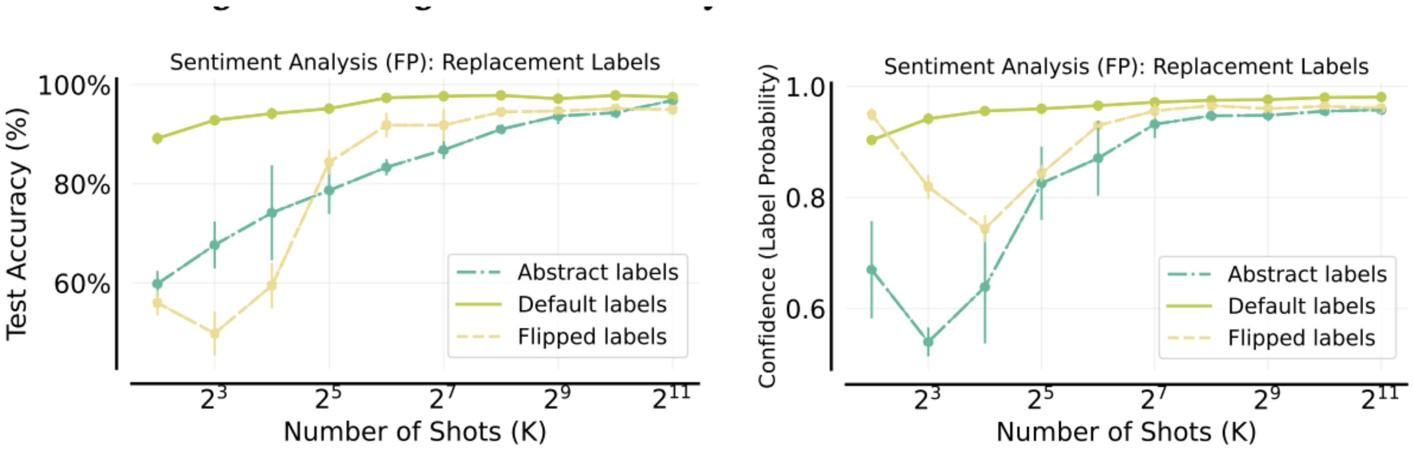
3. Analyzing Many-Shot ICL

- They also perform number of experiments analyzing the ICL in the paper and here is 2 of them

- Summary of results:
  - Next-token prediction loss may not be good predictor of ICL performance
  - Many-shot ICL can overcome pretraining biases, learn non-NLP prediction tasks, perform well compared to fine-tuning

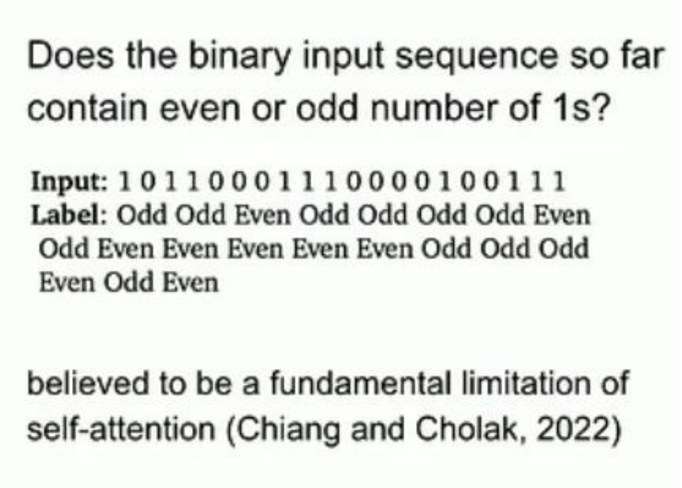
Overcoming Pre-Training Bias

- Kossen et al, 2023 suggests ICL has difficultly unlearning pre-training biases
- Figure 10: Sentiment analysis performance on flipped and abstract labels.
- Key point: Many-shot ICL can overcome pre-training biases, perform comparably to full fine-tuning



Many-Shot ICL for Non-NLP Tasks

- Learning high-dimensional functions (e.g., linear classification). This lets us do stress testing general applicability to possible unseen tasks .
- Outperforms a GPT-2 sized model trained from scratch on 20x more data

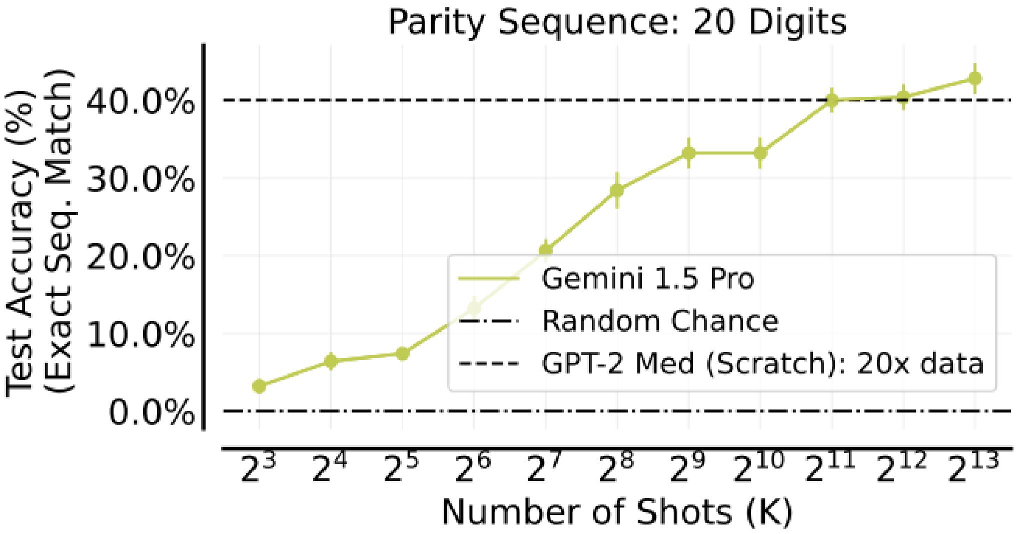


Does the binary input sequence so far contain even or odd number of 1s?

Input: 1 0 1 1 0 0 0 1 1 1 0 0 0 0 1 0 0 1 1 1
Label: Odd Odd Even Odd Odd Odd Odd Even Odd Even Even Even Even Even Odd Odd Odd Even Odd Even

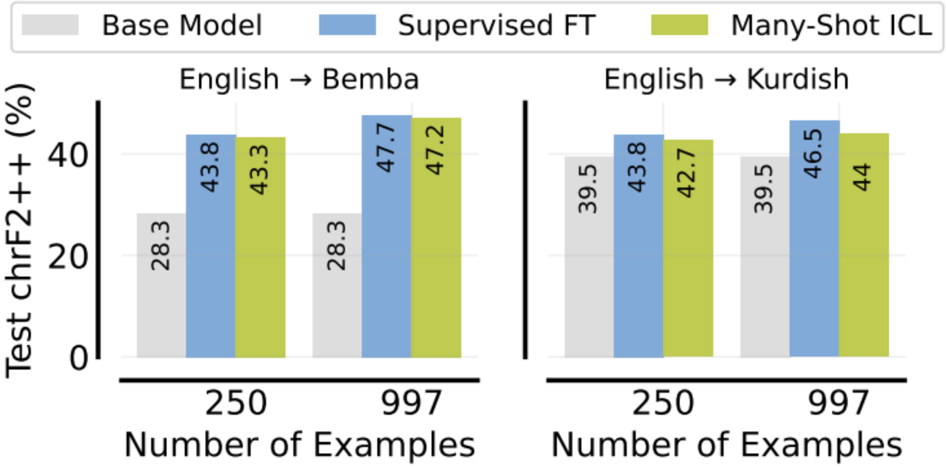
believed to be a fundamental limitation of self-attention (Chiang and Cholak, 2022)

t.



Comparison to Fine-Tuning

- Figure 13: Many-shot ICL vs. supervised fine-tuning for machine translation.
- Finding: Many-shot ICL can perform comparably to fine-tuning with fewer computational resources.



Why does the Many-Shot ICL work well?

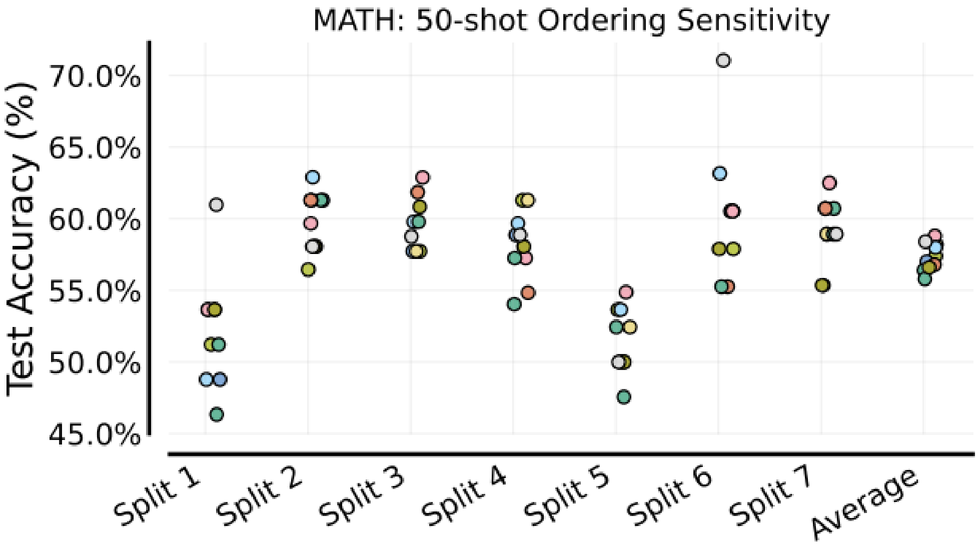
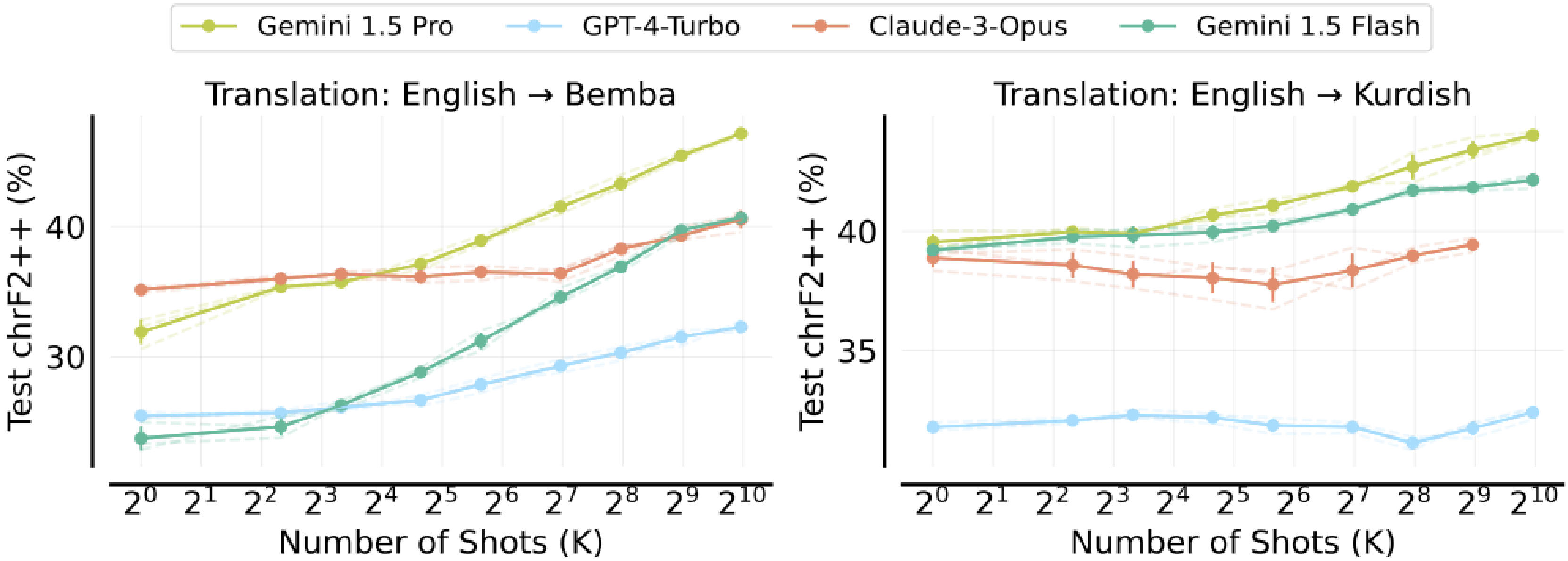


Figure 17 | **Many-Shot Sensitivity To Example Ordering.** Each colored data point represents a different random ordering of 50 in-context examples provided to Gemini 1.5 Pro.

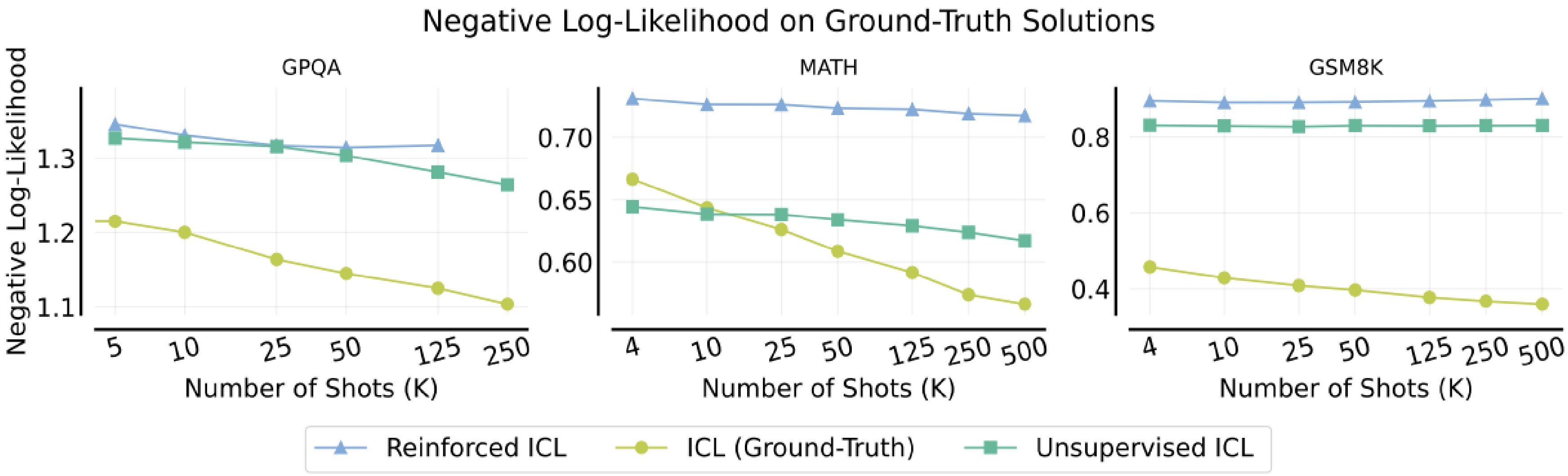
Many-shot ICL is affected by ordering of examples.

Model Comparision



Different LLM's exhibit varying degree of many-shot ICL capability

Long Scaling Laws do not predict ICL performance



NLL not reliable proxy when attempting to predict ICL performance for problem solving domains

Conclusion and Future Directions

- Summary: Many-shot ICL leads to performance gains across multiple tasks and reduces reliance on fine-tuning.
- Limitations: Ordering sensitivity and challenges with large example sets.
- Future work: Explore many-shot ICL capabilities across a wider range of models.

Questions?