Many-Shot In-Context Learning

NeurIPS, 2024 (Spotlight)

Rishabh Agarwal* , Avi Singh* , Lei M. Zhang† , Bernd Bohnet† , Luis Rosias† , Stephanie Chan† , Biao Zhang† , Ankesh Anand , Zaheer Abbas , Azade Nova , John D. Co-Reyes , Eric Chu , Feryal Behbahani , Aleksandra Faust and Hugo Larochelle

*Contributed equally, †Key contribution

Google DeepMind

Rishabh Agarwal

Avi Singh

Lei Zhang

Bernd Bohnet

L

Luis Rosias

Stephanie Chan

B

Biao Zhang

Ankesh Anand

Zaheer Abbas

Azade Nova

JD Co-Reyes

Eric Chu

Feryal Behbahani

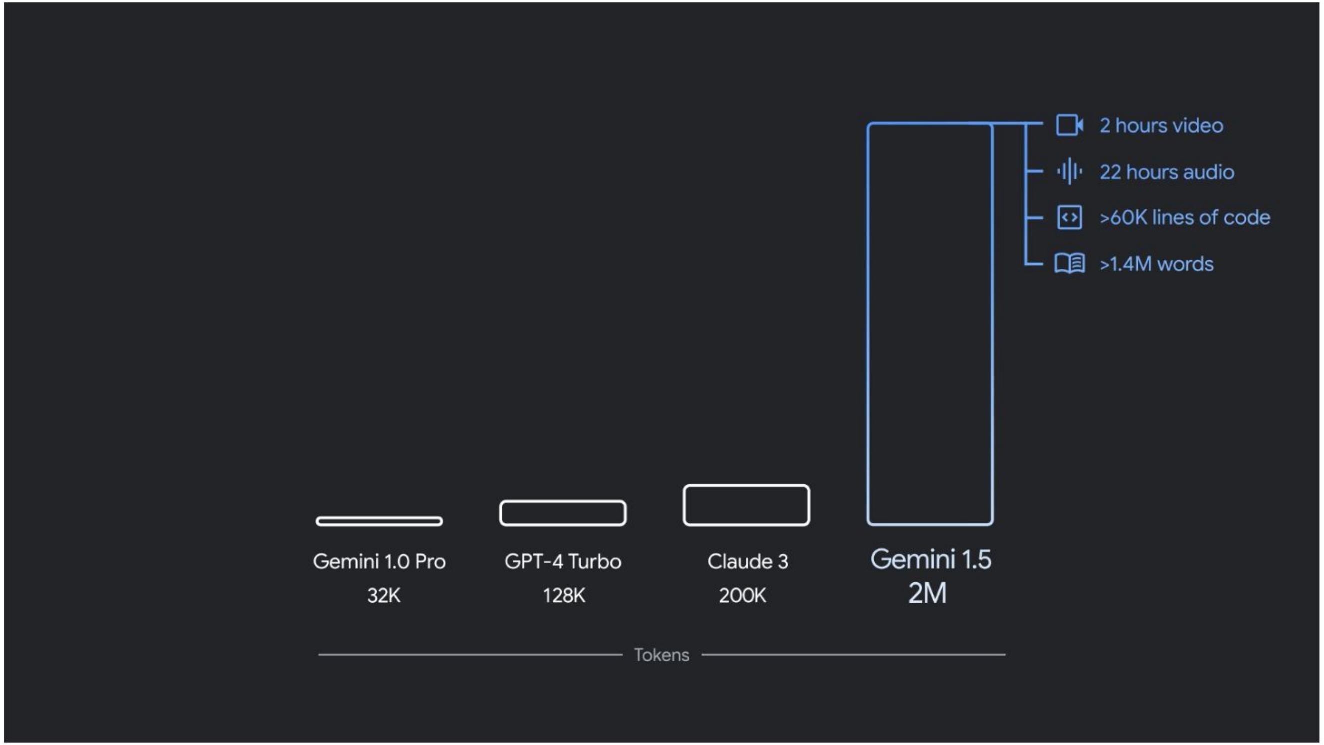
Aleksandra Faust

Hugo Larochelle

From few-shot to many-shot in-context learning (ICL)

1

From few-shot to many-shot ICL

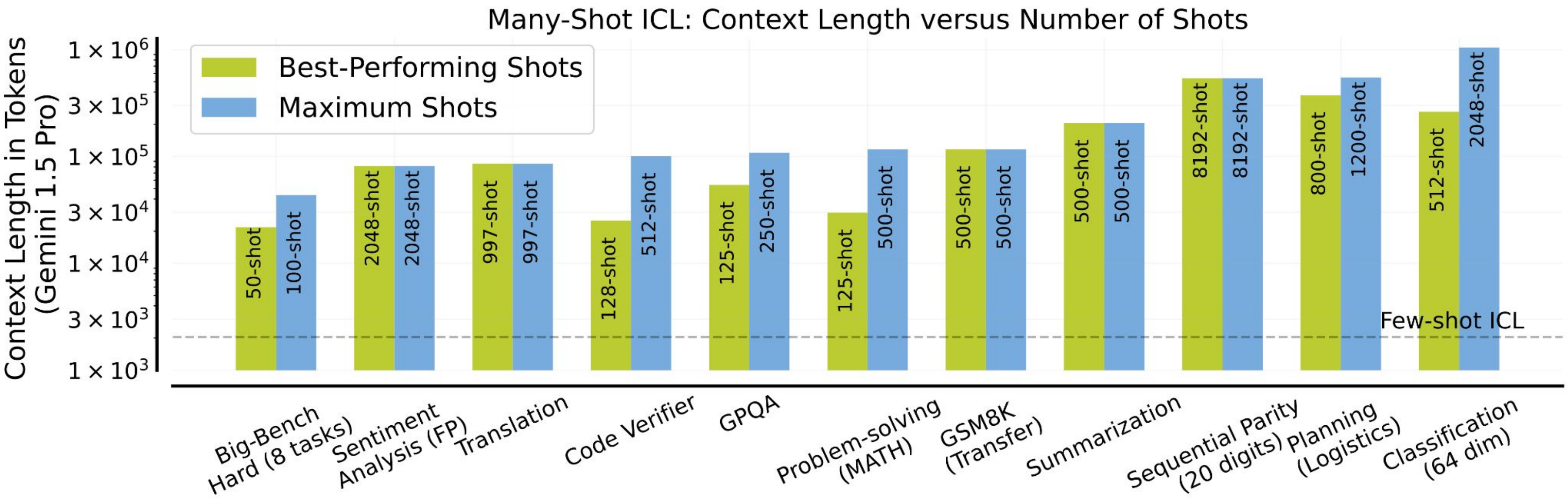


Context lengths of leading foundation models compared with Gemini 1.5's 2 million token capability

Image from Google IO Gemini blog post

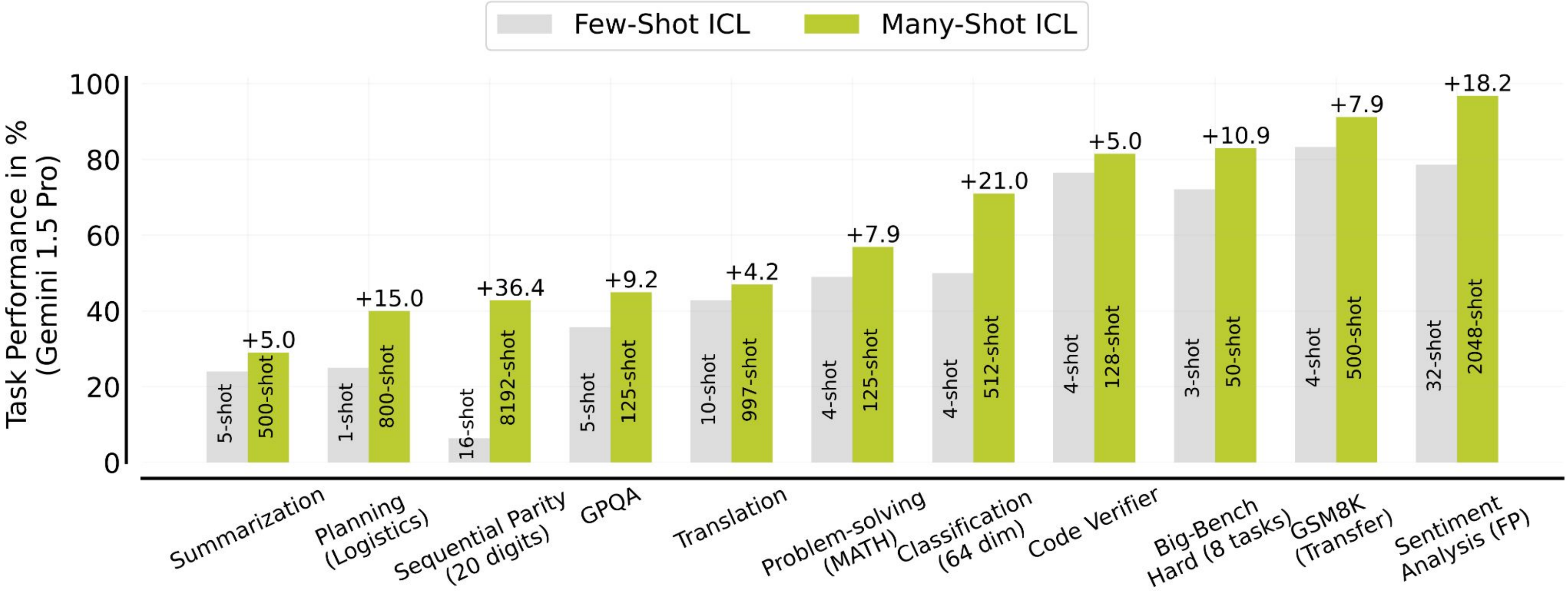
From few-shot to many-shot ICL

How many shots is “many-shot”?



From few-shot to many-shot ICL

Does many-shot ICL improve performance? Yes!

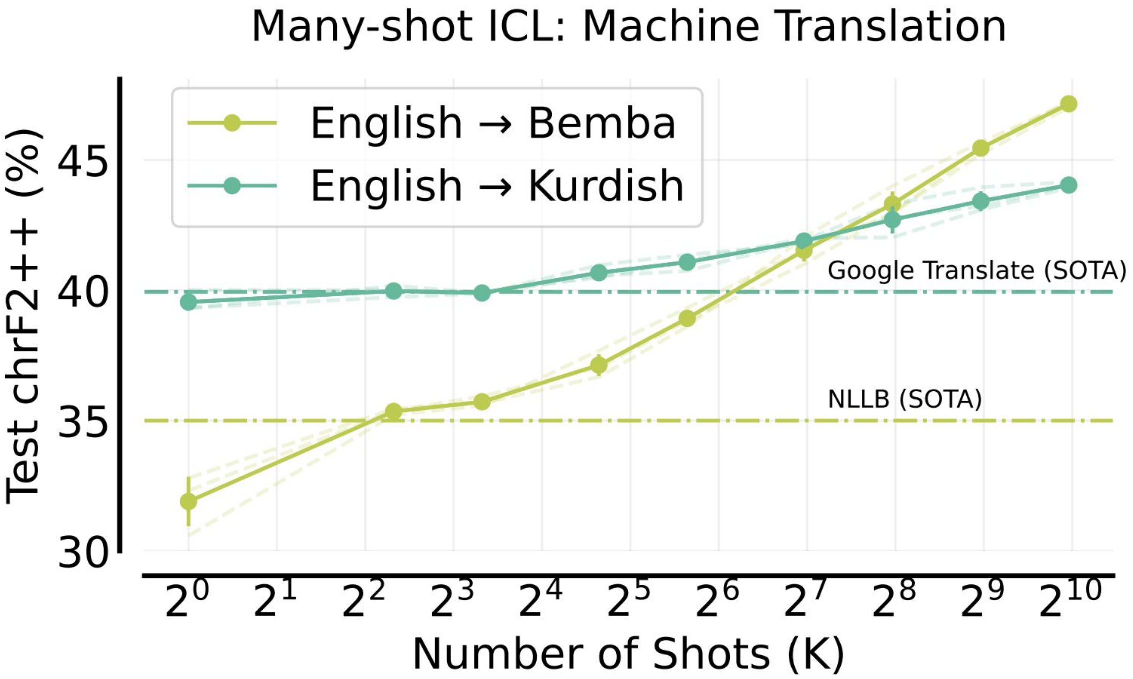


Many-shot ICL examples

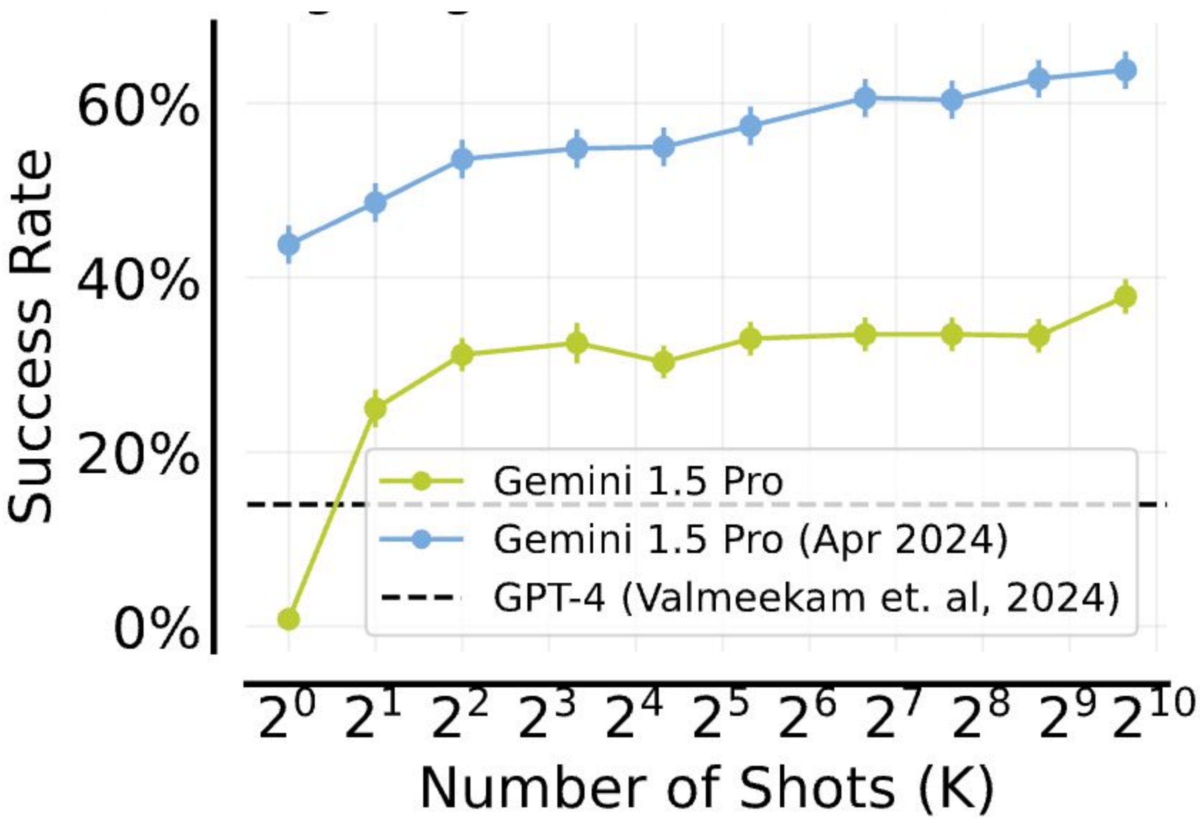
2

Machine translation on low-resource languages

Beating SOTA systems using many-shot ICL.

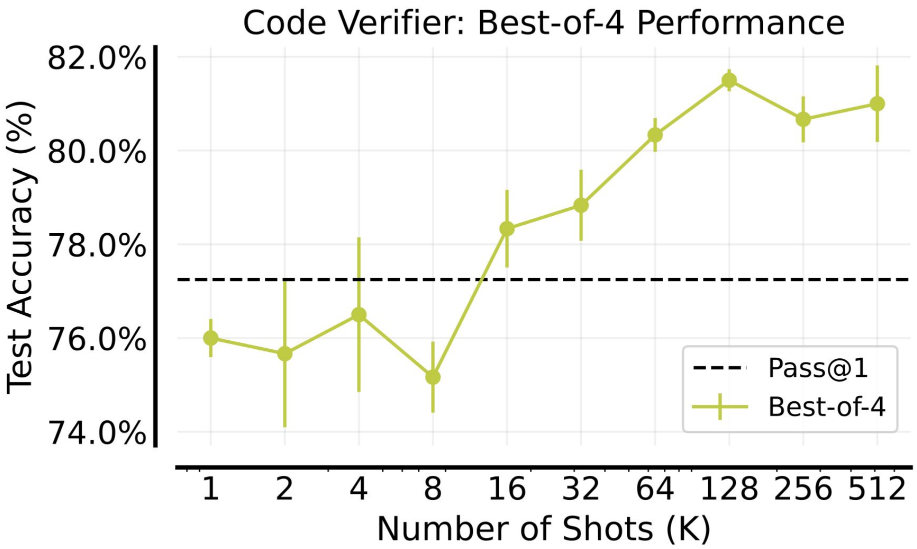


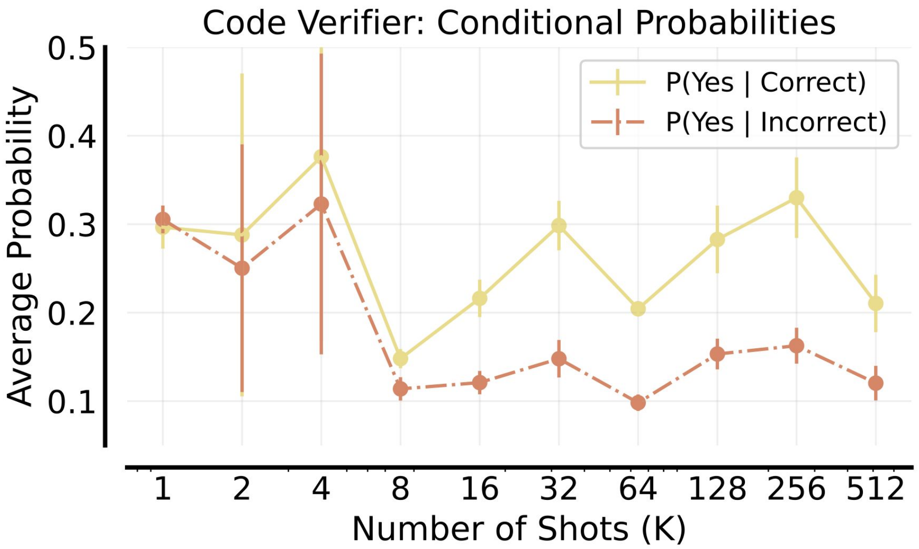
Logistics Planning



Code verifier

Code verifier without fine-tuning!





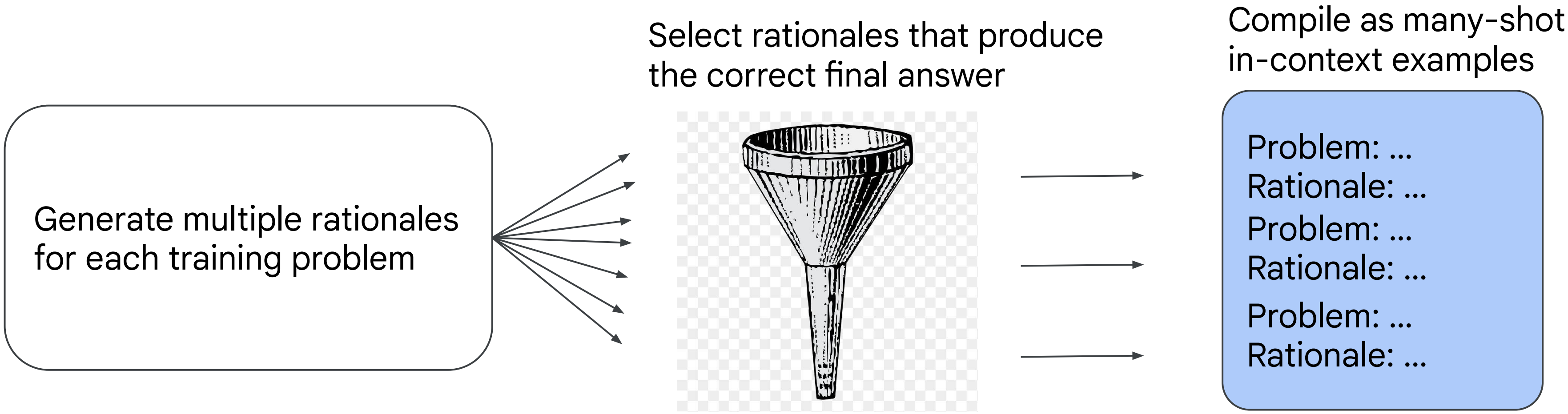
Many more examples of effective many-shot ICL:
e.g., Planning, Summarization. See paper for details!

Many-shot learning without human-written rationales

3

Human-written rationales or demonstrations can be expensive to collect… can we do without?

Reinforced ICL: use model-generated rationales



Generate multiple rationales for each training problem

Select rationales that produce the correct final answer

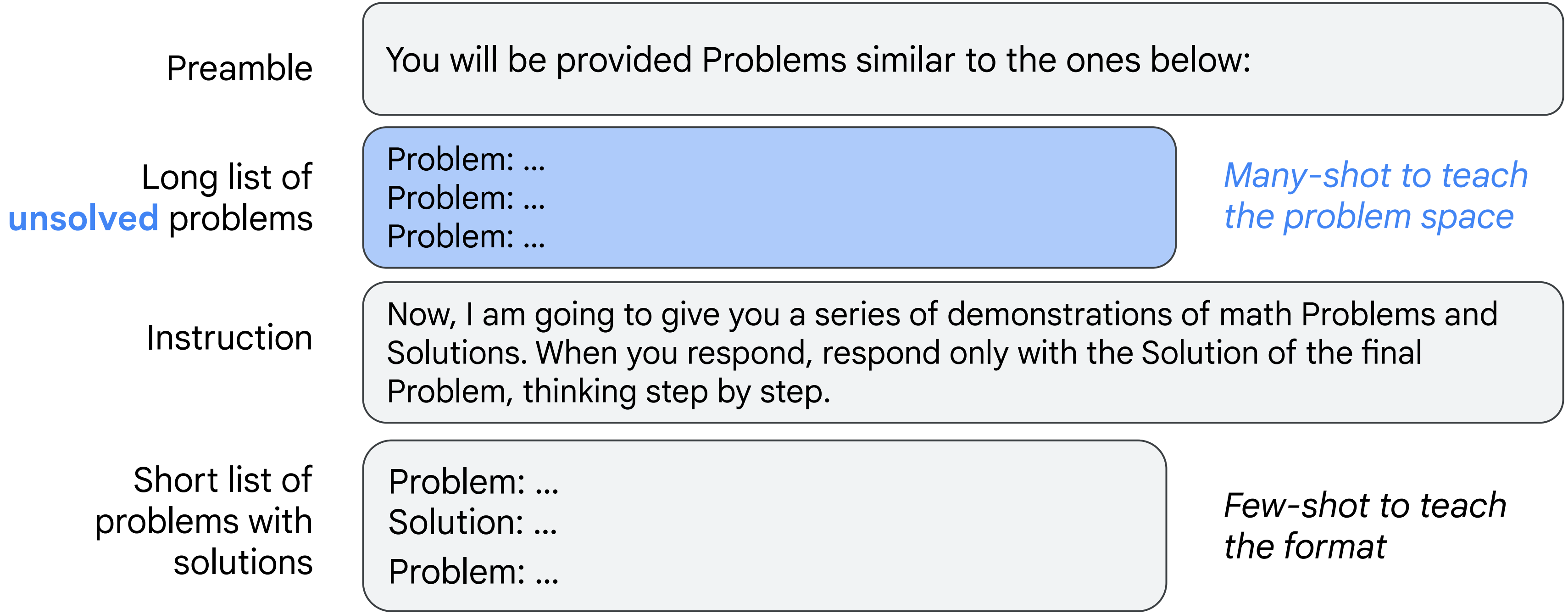
Compile as many-shot in-context examples

Problem: …
Rationale: …
Problem: …
Rationale: …
Problem: …
Rationale: …

*called "reinforced" because of equivalence to expectation-maximization RL algorithm

Human-written rationales or demonstrations can be expensive to collect… can we do without?

Unsupervised ICL: get rid of rationales/solutions entirely!



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

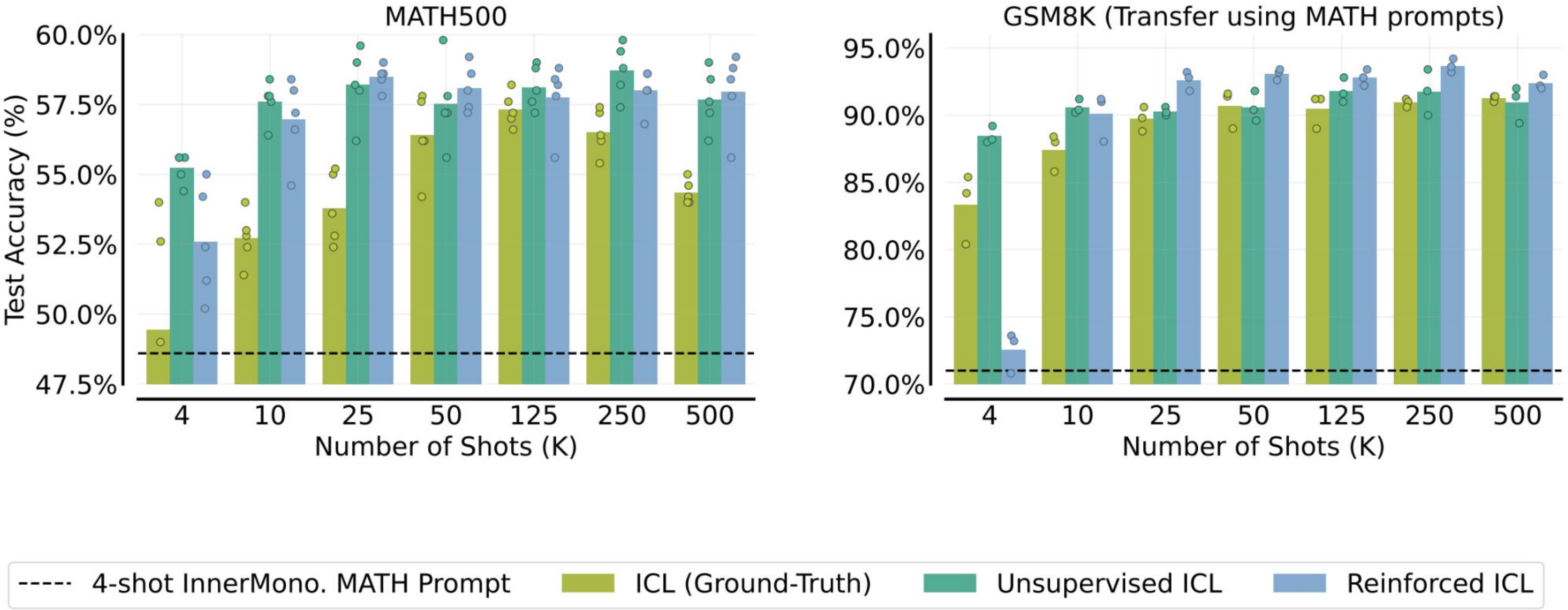
Problem: …
Solution: …
Problem: …

Few-shot to teach the format

Problem-Solving: Hendrycks MATH & GSM8K

Reinforced and Unsupervised ICL can outperform ICL with human-written solutions!

The Hendrycks MATH prompts transfer well to GSM8k (another math dataset)

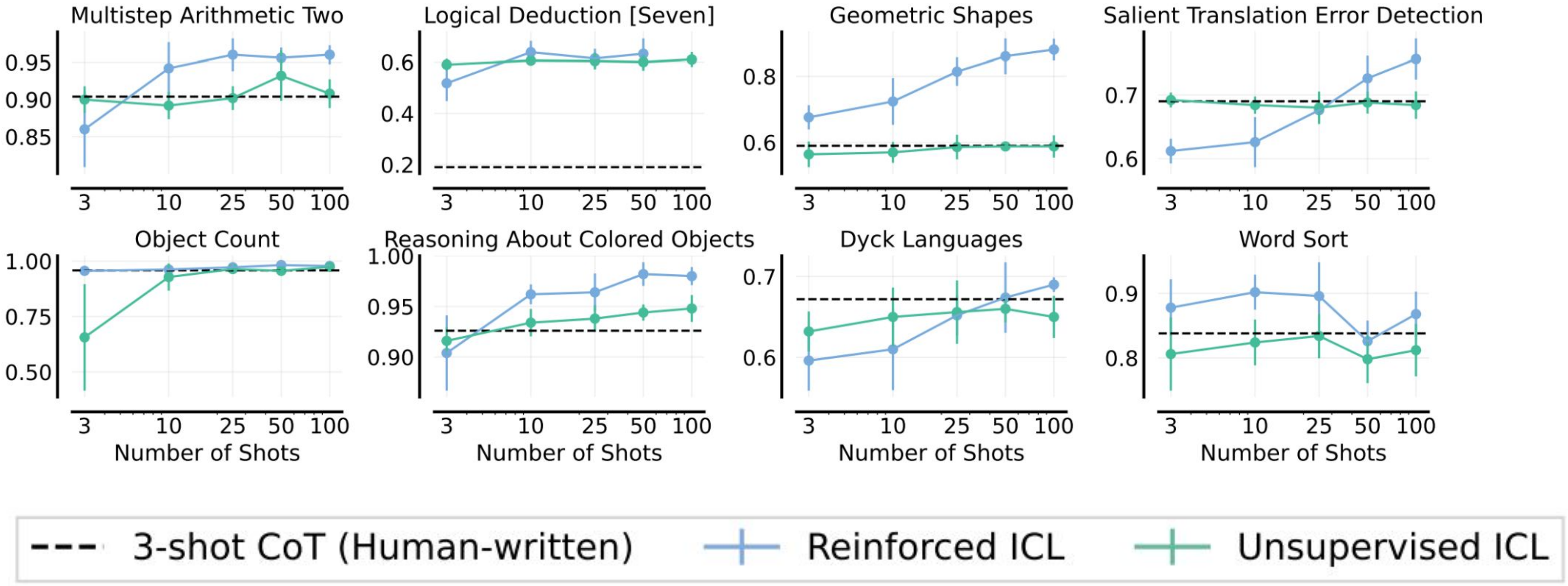


(human-written)

Algorithmic and Symbolic Reasoning: Big-Bench Hard

Generally: Reinforced > Unsupervised > Human-written

- with greater improvement for more shots2

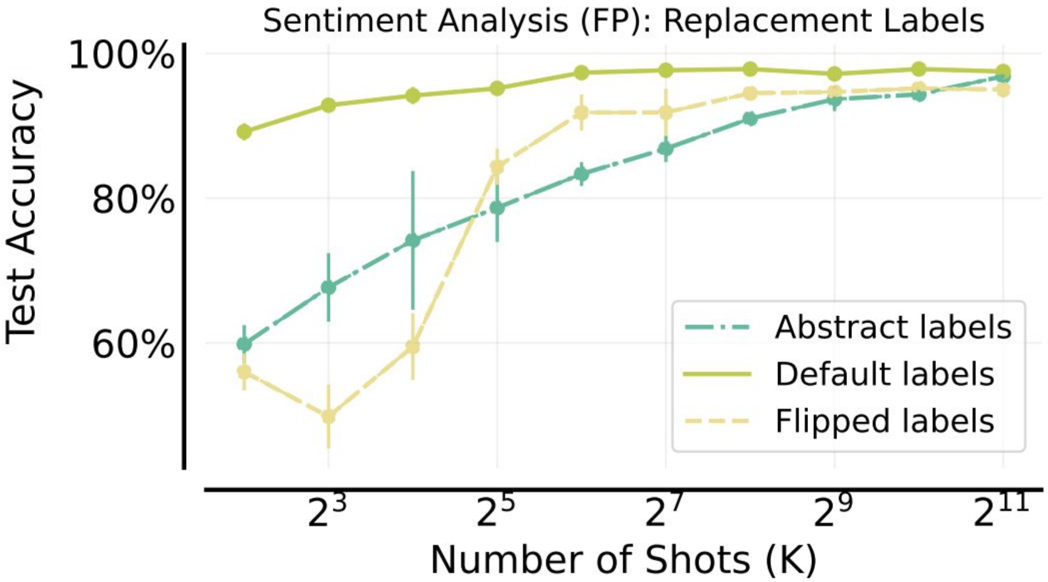


Analyzing many-shot ICL

4

Many-shot ICL can overcome pre-training biases

Previous work (Kossen et al, 2023) suggest that ICL has difficulty unlearning biases derived from pre-training data…

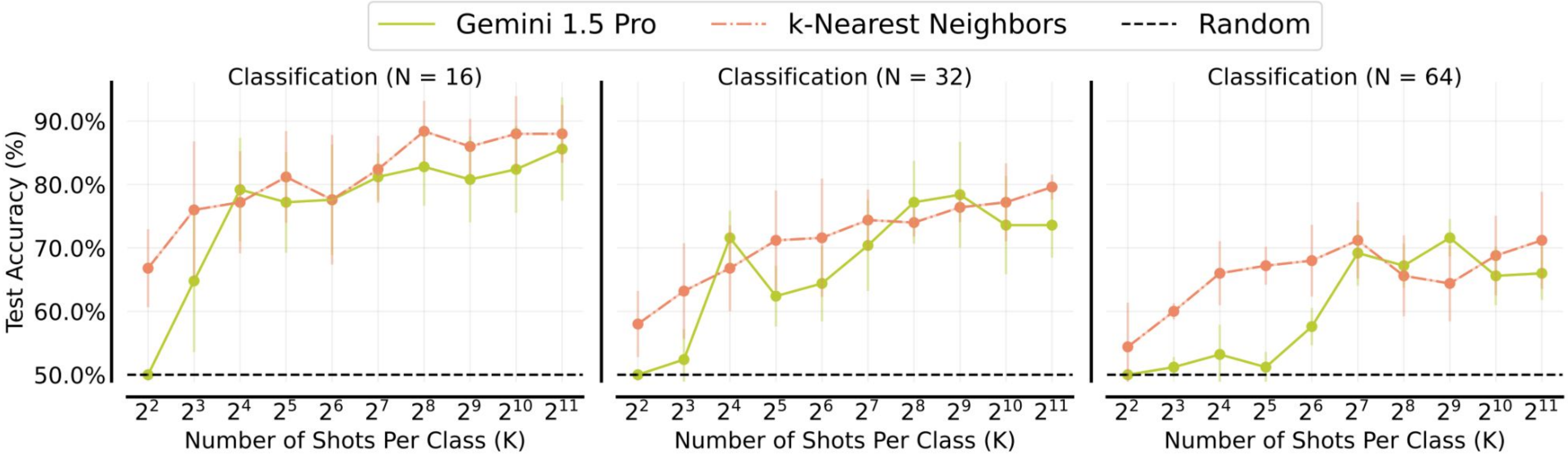


Default (original)	Flipped (rotated)	Abstract
negative	neutral	A
neutral	positive	B
positive	negative	C

…but with enough shots, new labels eventually approach performance of original labels

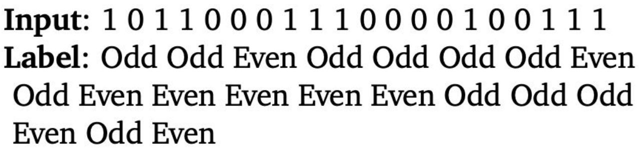
High-dimensional functions:
Binary Linear Classification in High Dimensions

Many-shot ICL nearly matches strong baseline (k-nearest neighbors)



High-dimensional functions:
Sequential Parity (20 digits)

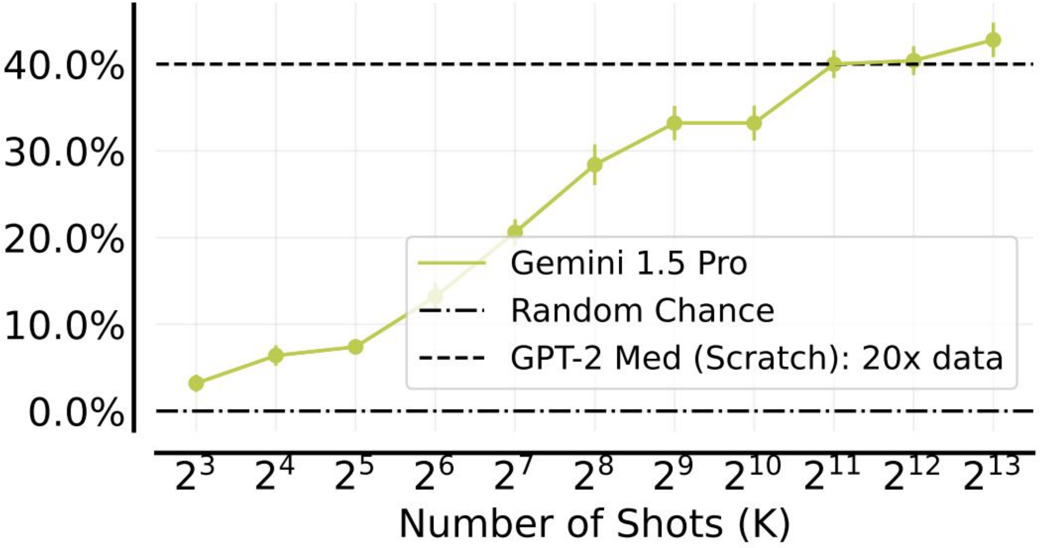
Does the binary input sequence so far contain even or odd number of 1s?



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

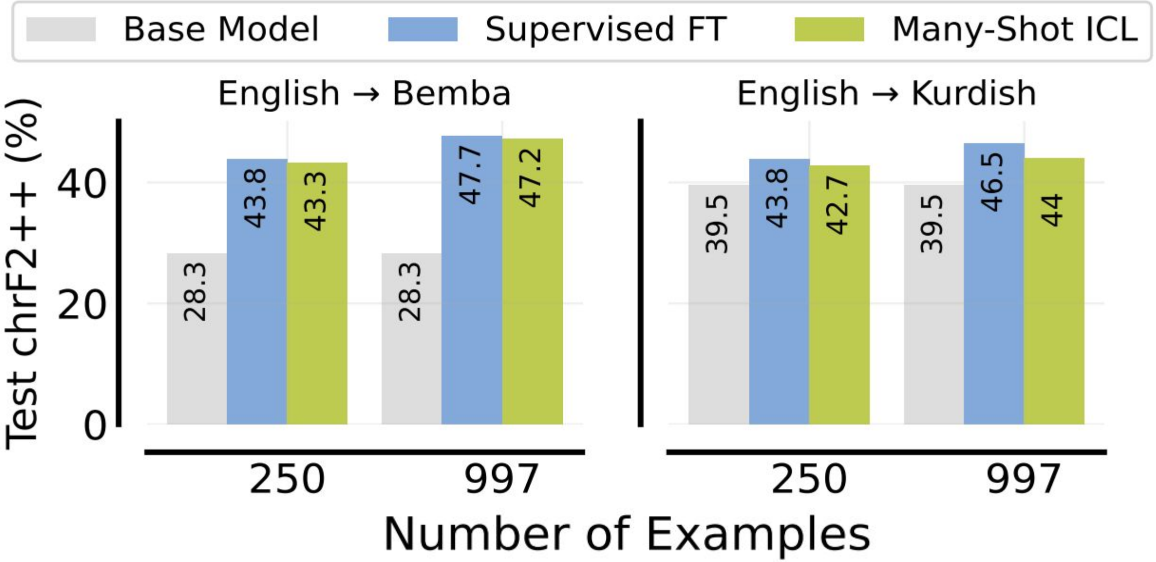
believed to be a fundamental limitation of self-attention (Chiang and Cholak, 2022)

Many-shot ICL improves monotonically until 2^13 examples!



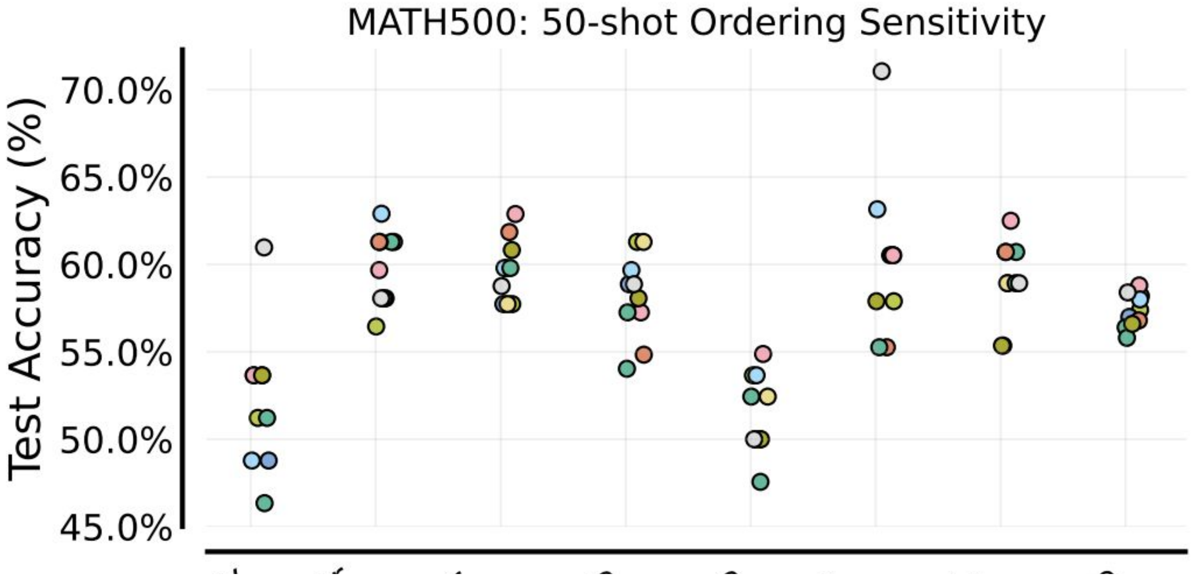
And outperforms a GPT-2 sized model trained from scratch on 20x more data

Many-shot ICL can have similar performance to SFT (translation task)



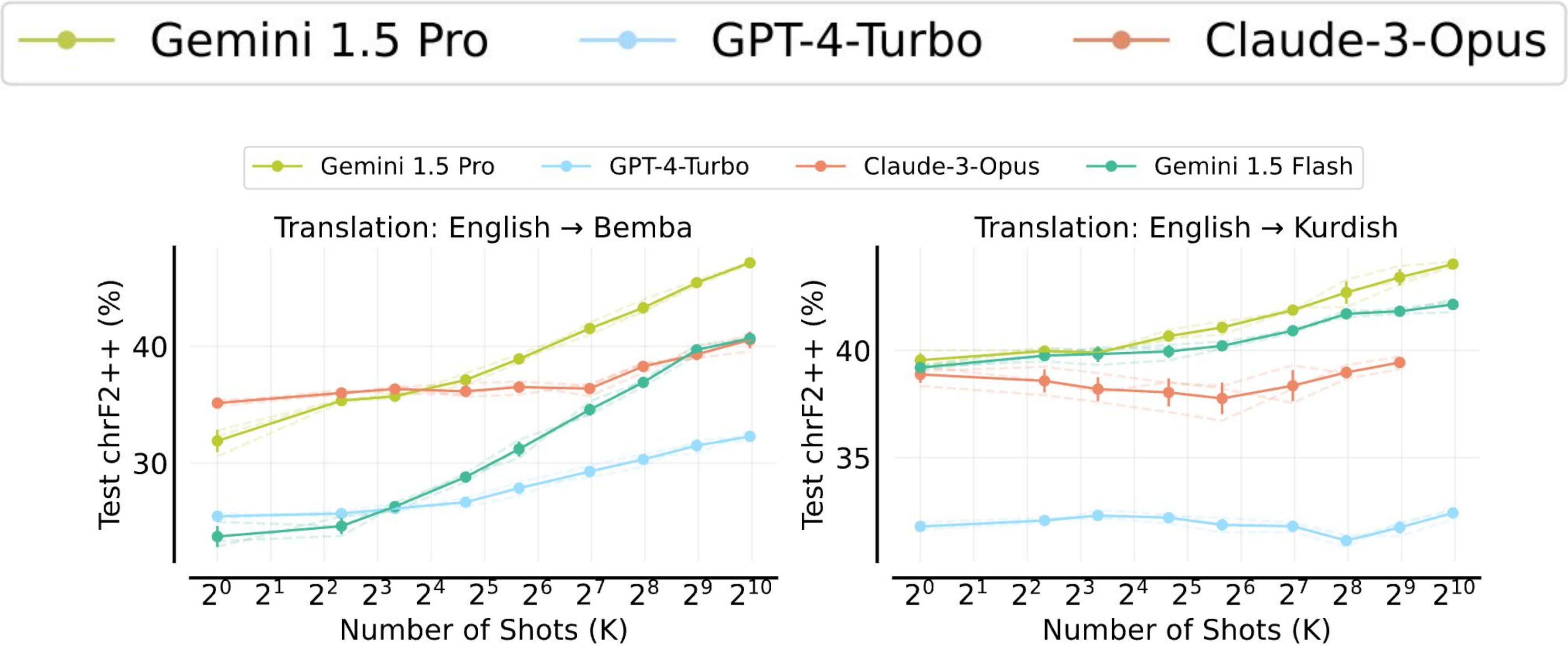
- ICL has no training cost but potentially higher inference cost (can mitigate with context caching)

Many-shot ICL can be sensitive to example ordering



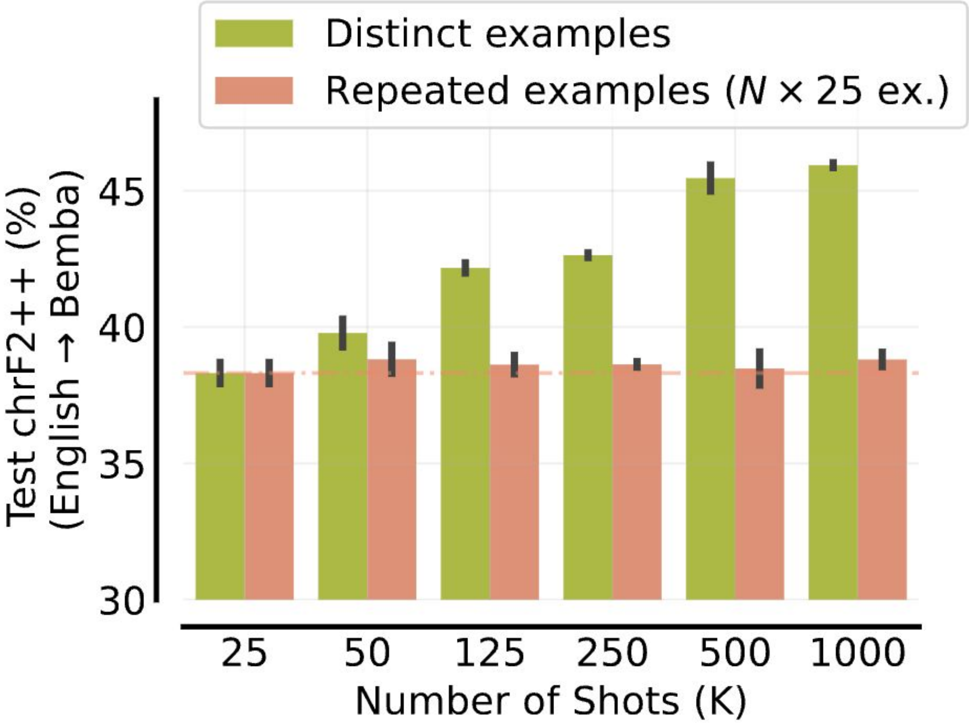
Each colored point is a different ordering

Comparison of frontier models



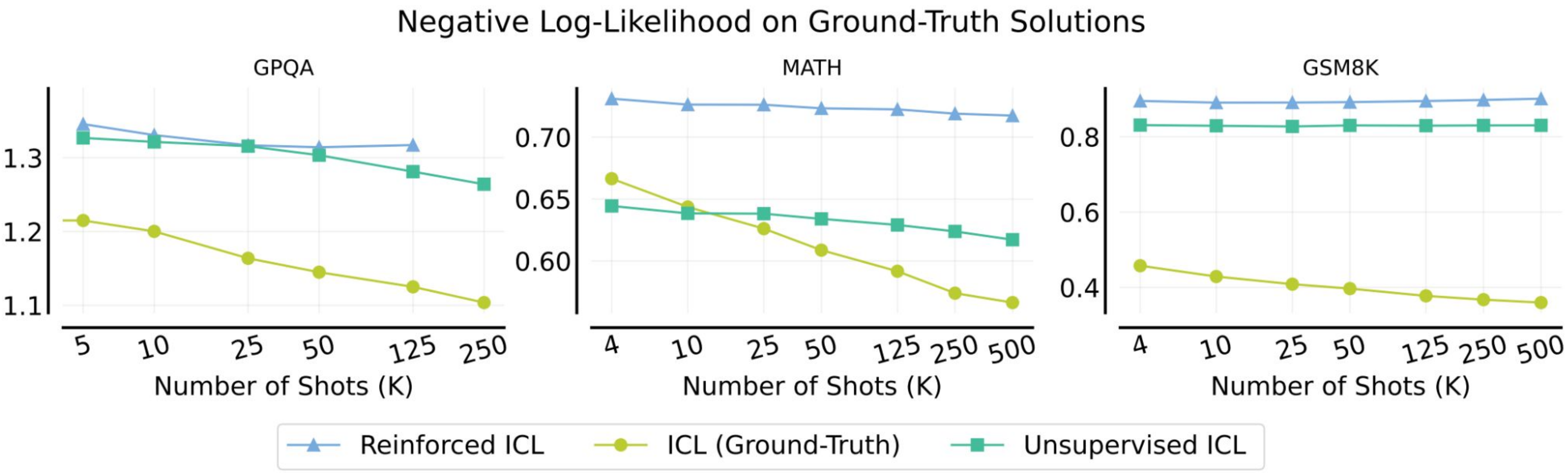
Varying levels of benefit from many-shot ICL

Increasing Context Length or More Information?



Many-shot performance with distinct examples vs repeating the same 25 examples N times on low-resource MT.

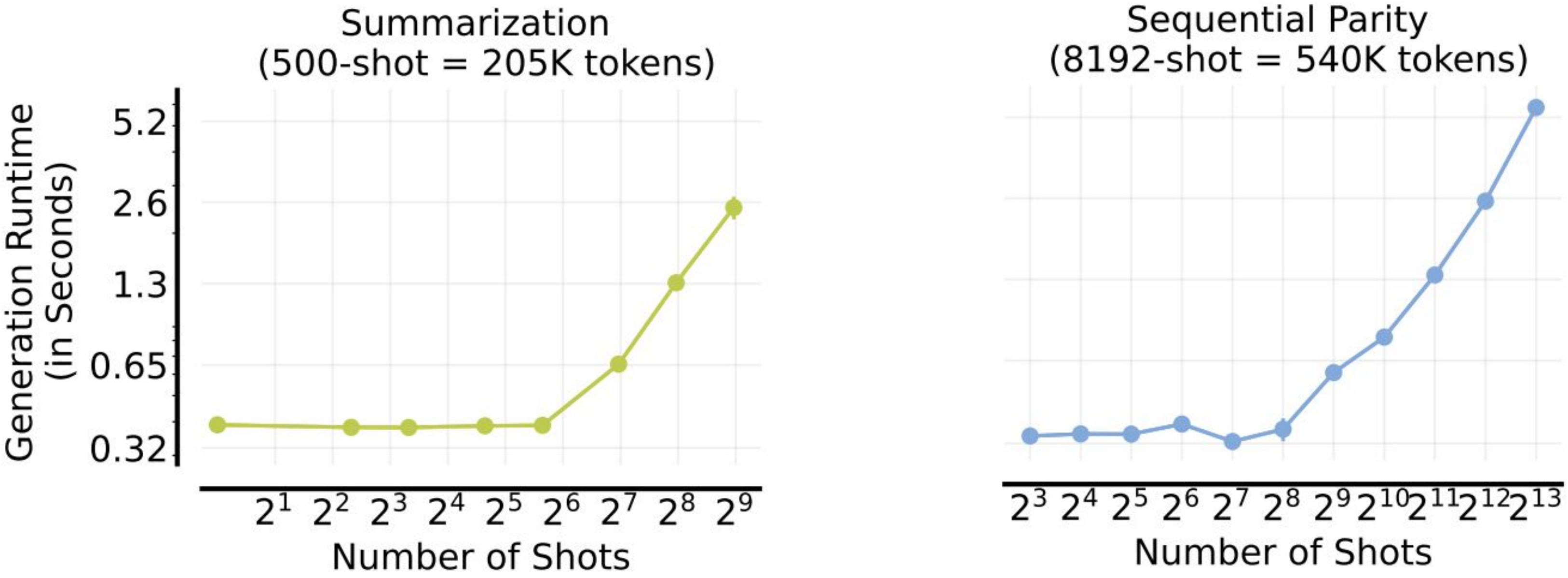
Long-context scaling laws may not predict ICL performance



NLL is not a reliable predictor of ICL performance

- NLL consistently decreases, even though ICL worsens beyond 125 shots
- NLL for human-written rationales is lower than for model-written rationales, even though actual performance is often worse

Inference costs



Summary

1. **Many-shots can improve performance up to 1000s of shots**
   Long-context models enable this (to varying degrees)

2. **Model-generated or unsupervised prompts** can often outperform human-written prompts

3. **Analyses**:
   - Many-shot ICL can overcome pretraining biases
   - Many-shot ICL can have similar performance to SFT
   - NLL is not a reliable predictor of ICL performance

Q: What are the mechanisms underlying many-shot learning? Why do particular tasks benefit more?

Q: Why does performance sometimes degrade after many shots?

Q: Why does example ordering matter?