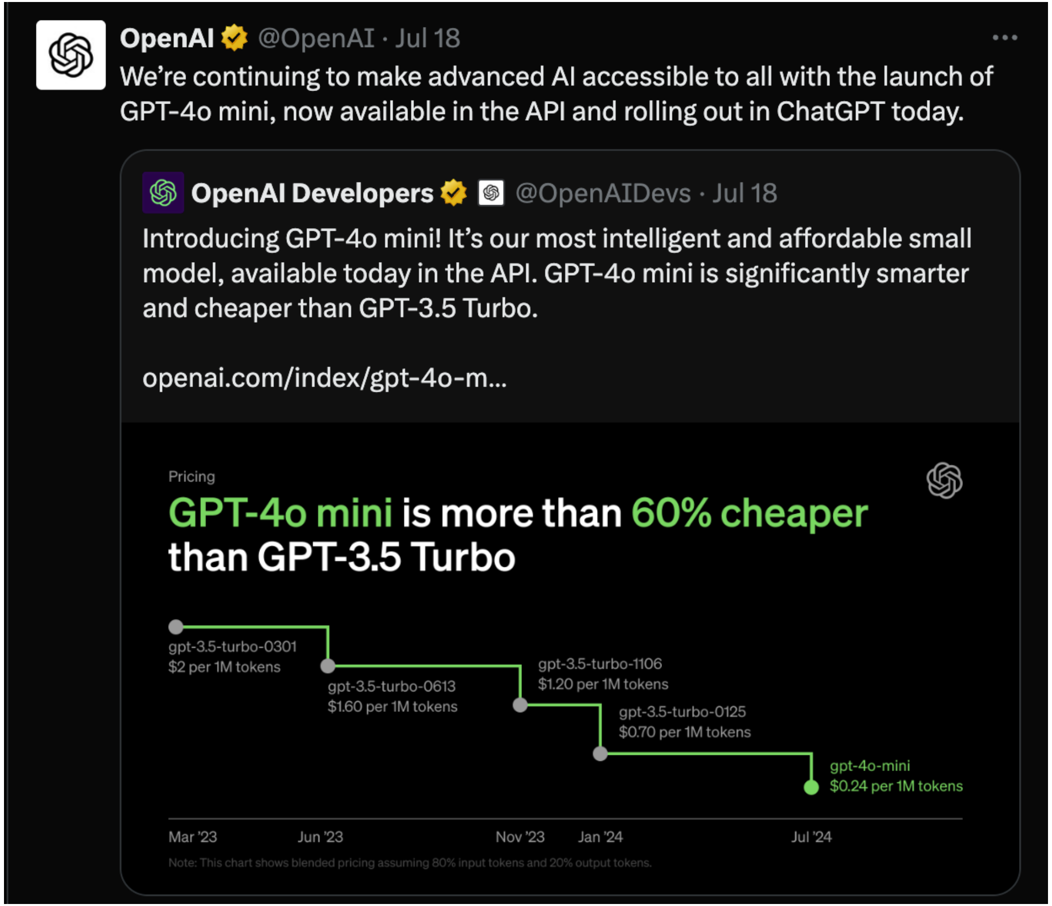
MODEL COMPRESSION: QUANTIZATION AND KNOWLEDGE DISTILLATION

MOHAMAD KHAJEZADE

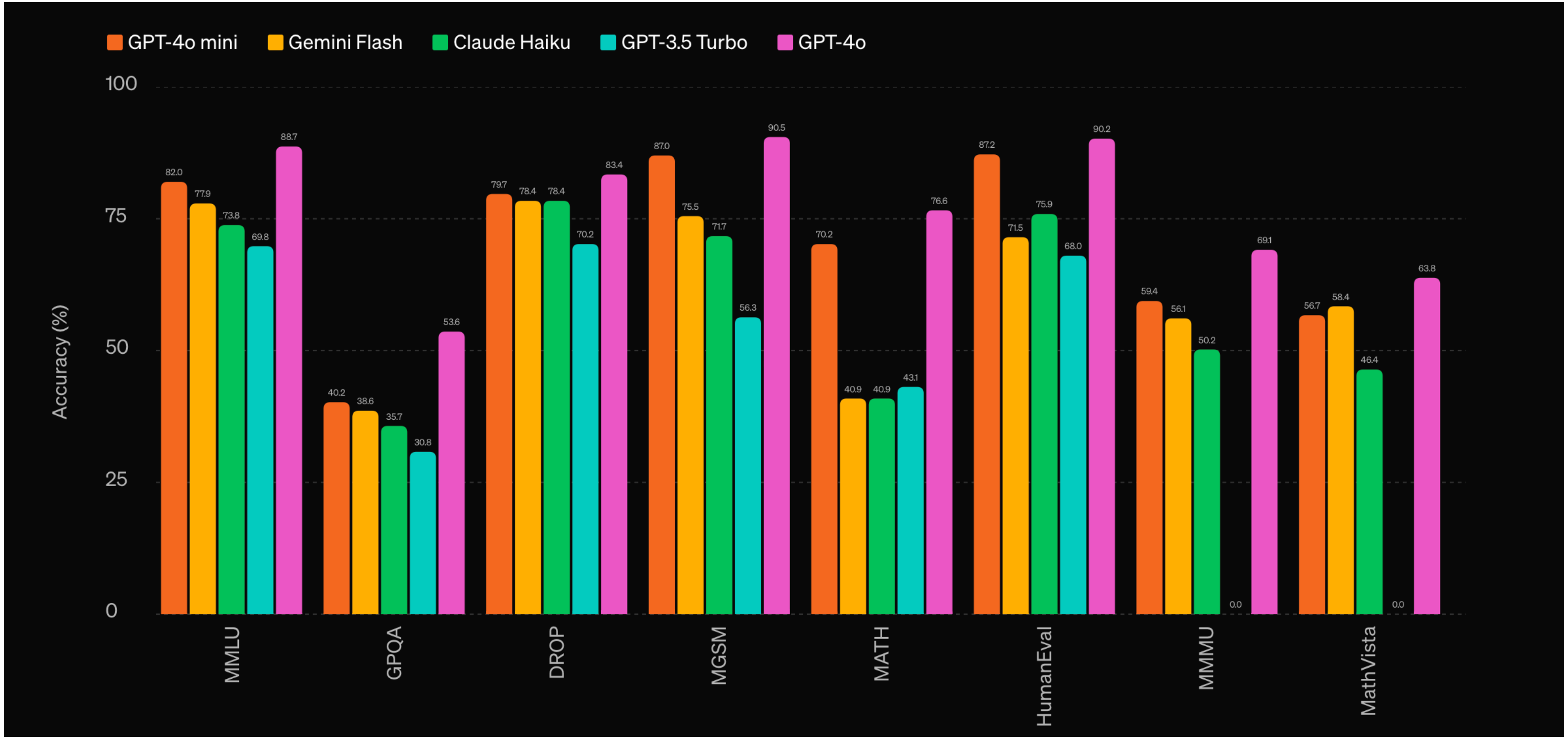
SUMMER, 2024

IMPORTANCE OF SMALL MODELS



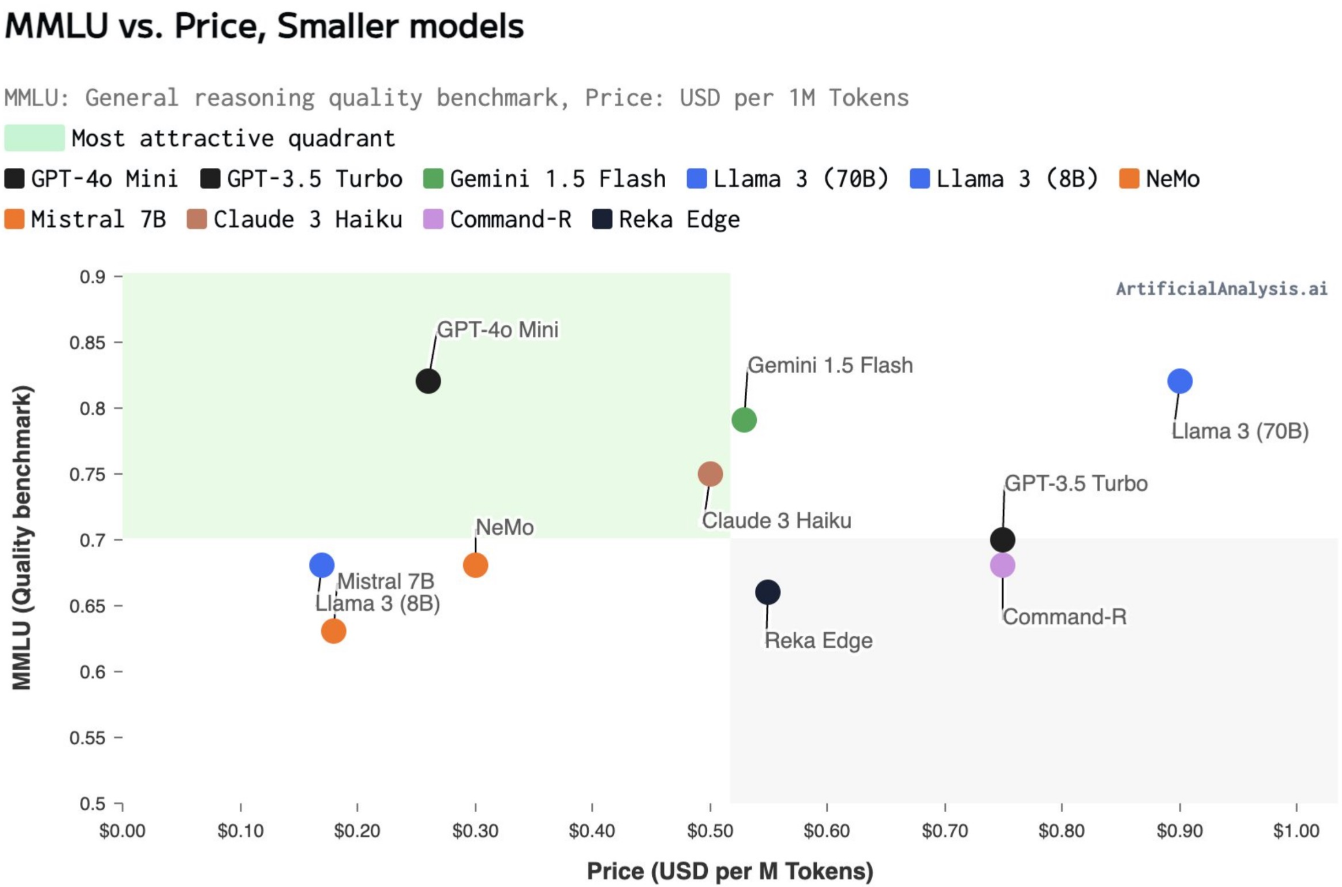
[OpenAI GPT-4o mini anouncement]

IMPORTANCE OF SMALL MODELS



[GPT-4o mini evaluations]

IMPORTANCE OF SMALL MODELS



IMPORTANCE OF SMALL MODELS

Andrej Karpathy
@karpathy

LLM model size competition is intensifying… backwards!

My bet is that we'll see models that "think" very well and reliably that are very very small. There is most likely a setting even of GPT-2 parameters for which most people will consider GPT-2 "smart". The reason current models are so large is because we're still being very wasteful during training - we're asking them to memorize the internet and, remarkably, they do and can e.g. recite SHA hashes of common numbers, or recall really esoteric facts. (Actually LLMs are really good at memorization, qualitatively a lot better than humans, sometimes needing just a single update to remember a lot of detail for a long time). But imagine if you were going to be tested, closed book, on reciting arbitrary passages of the internet given the first few words. This is the standard (pre)training objective for models today. The reason doing better is hard is because demonstrations of thinking are "entangled" with knowledge, in the training data.

Therefore, the models have to first get larger before they can get smaller, because we need their (automated) help to refactor and mold the training data into ideal, synthetic formats.

It's a staircase of improvement - of one model helping to generate the training data for next, until we're left with "perfect training set". When you train GPT-2 on it, it will be a really strong / smart model by today's standards. Maybe the MMLU will be a bit lower because it won't remember all of its chemistry perfectly. Maybe it needs to look something up once in a while to make sure.

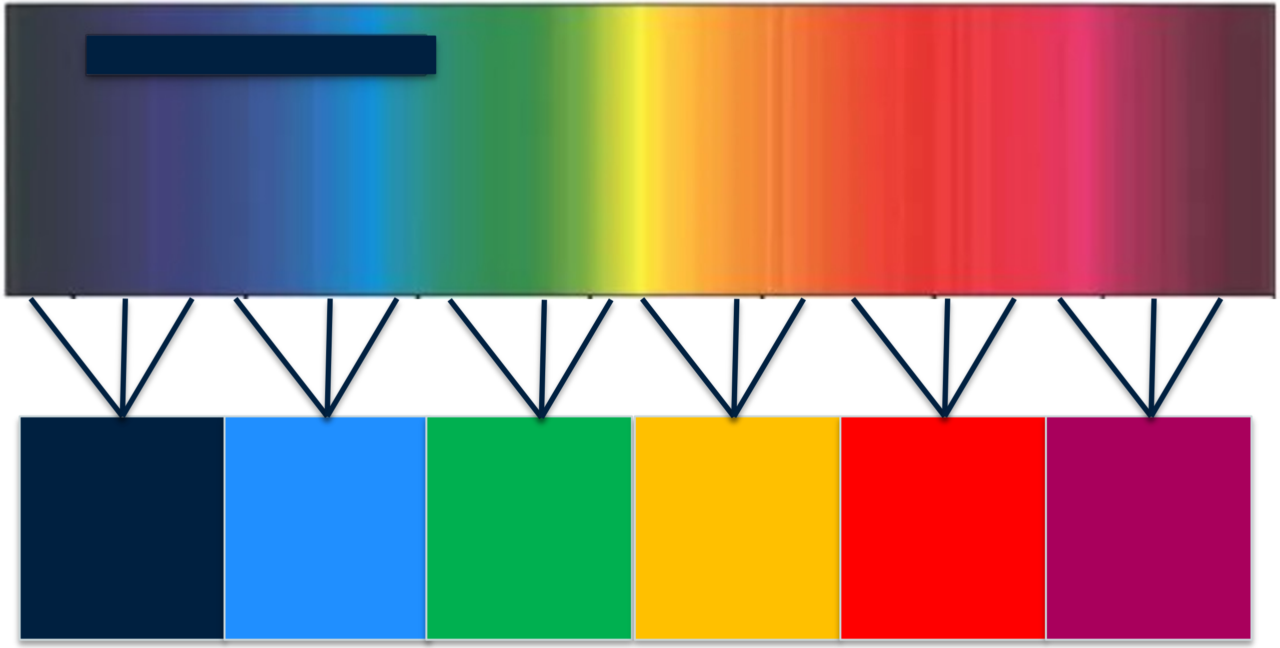
Andrej's Tweet

SUMMARIZING ANDREJ’S TWEET

- A competition has started, which contrasts with the previous competition: to create smaller models
- Creating large models before small models was inevitable and an essential step.
- Models were large because their training was not efficient. However, they have absorbed the knowledge of the internet.
- Now, it is possible to effectively use this knowledge to create smaller models.

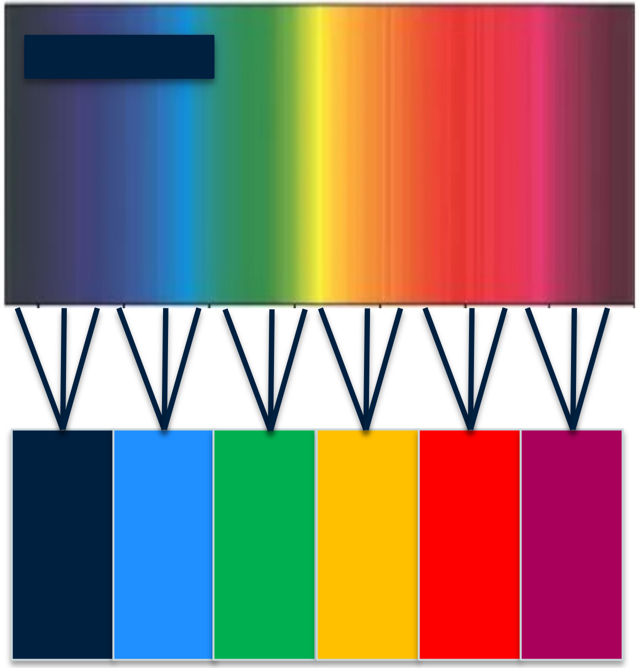
QUANTIZATION

LARGE SET OF POSSIBLE VALUES



SMALL SET OF POSSIBLE VALUES

QUANTIZATION: LARGE LANGUAGE MODELS



- **Often means reducing the precision of the weights' values:**
  - ❖ Smaller precision results in smaller memory requirement
  - ❖ Smaller precision results in faster inference

QUANTIZATION: LARGE LANGUAGE MODELS

0.34	**3.75**	**5.64**
1.12	2.7	-0.9
-4.7	0.68	1.43

FP 32

→

8	**84**	**126**
25	61	-20
-106	15	32

INT 8

QUANTIZATION: LARGE LANGUAGE MODELS

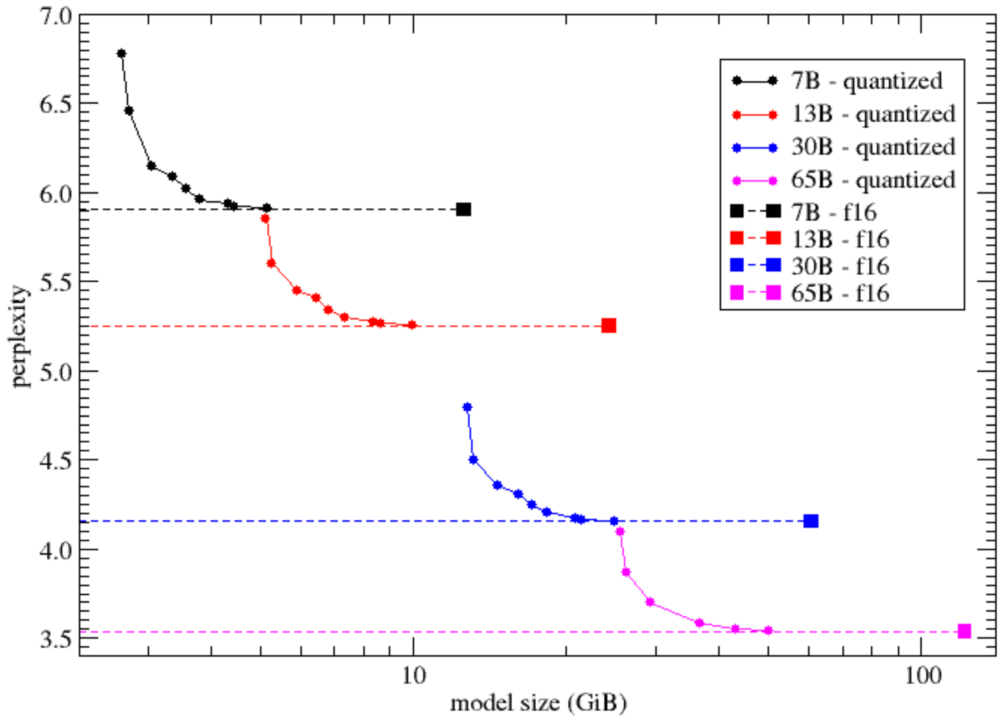
Precision	Range	Possible Values
FP 32	$-3.4 \times 10^{38} \rightarrow 3.4 \times 10^{38}$	4.2×10^{9}
INT 8	$-128 \rightarrow 127$	256

QUANTIZATION: LARGE LANGUAGE MODELS

$$\mathbf{X}_{\text{quant}} = \text{round}\left(\frac{127}{\max|\mathbf{X}|} \cdot \mathbf{X}\right)$$

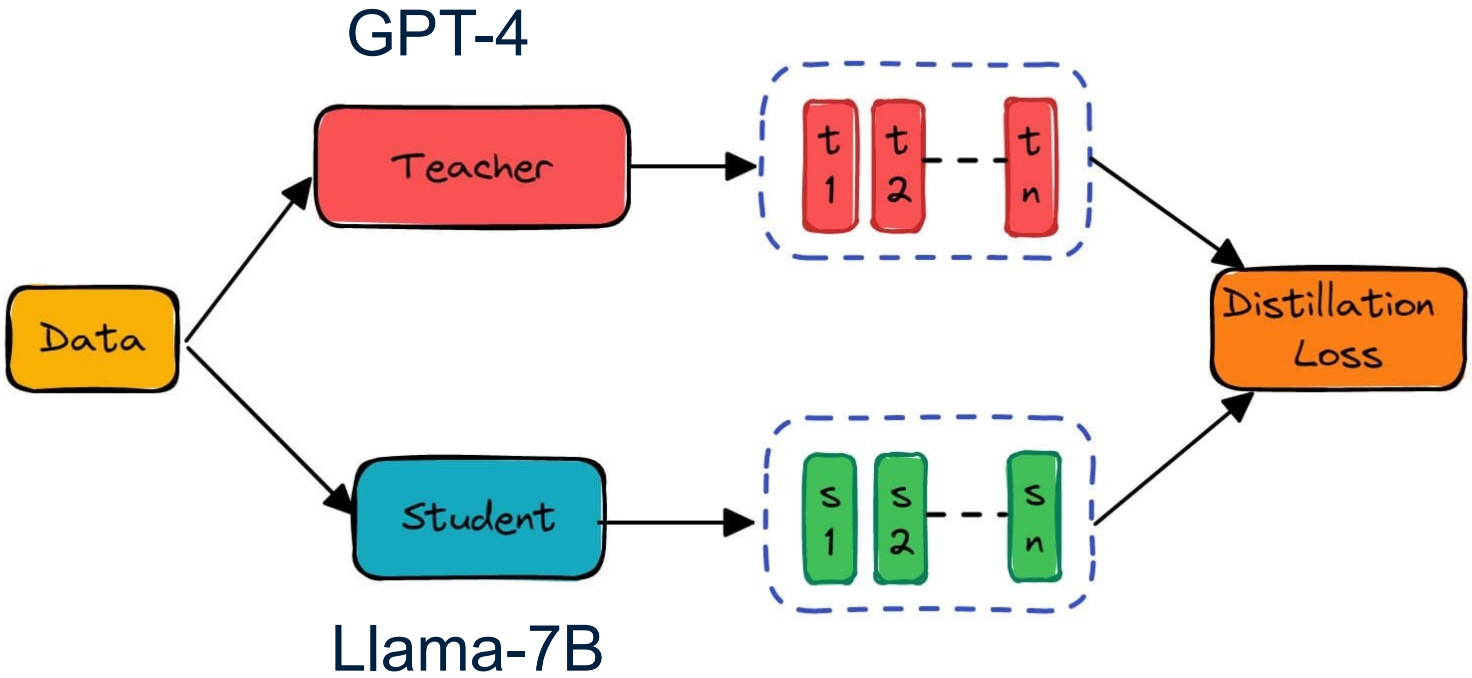
$$\mathbf{X}_{\text{dequant}} = \frac{\max|\mathbf{X}|}{127} \cdot \mathbf{X}_{\text{quant}}$$

QUANTIZATION: LARGE LANGUAGE MODELS



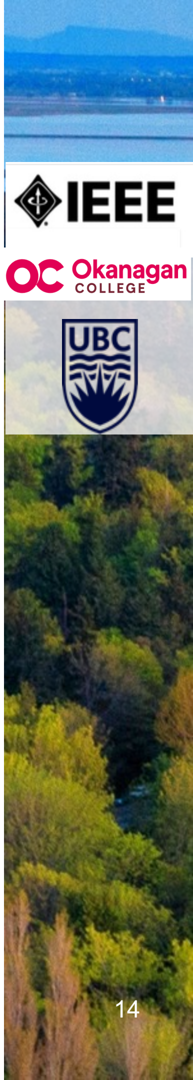
[Qualitative comparison]

INTRODUCTION

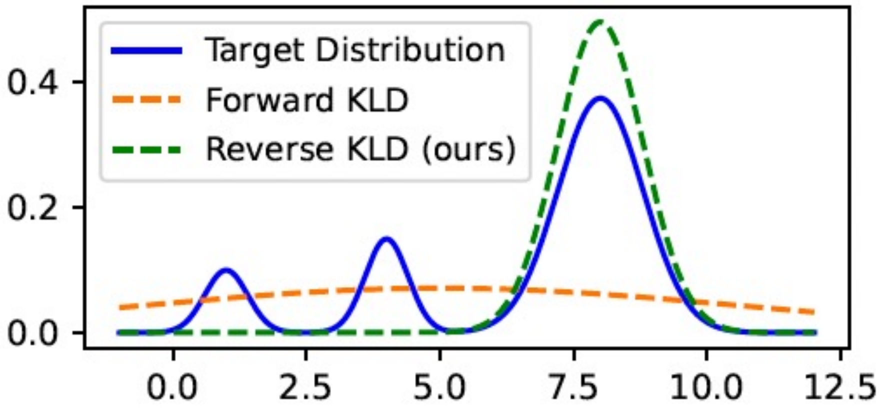


INTRODUCTION

- black-box KD and white-box KD
- black-box KD has shown promising results in fine-tuning small models
- white-box KD approaches are mostly studied for small (<1B parameters) language understanding models



INTRODUCTION

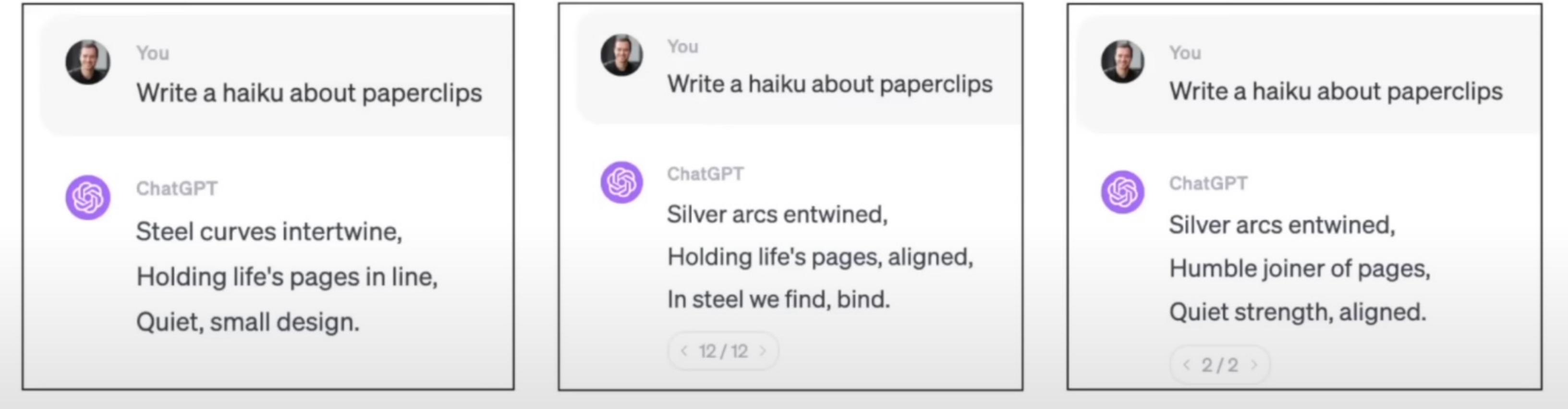


Problem: Student doesn't have the capacity of the teacher

Solution: Using the teacher output as a signal to improve student performance

[MiniLLM paper]

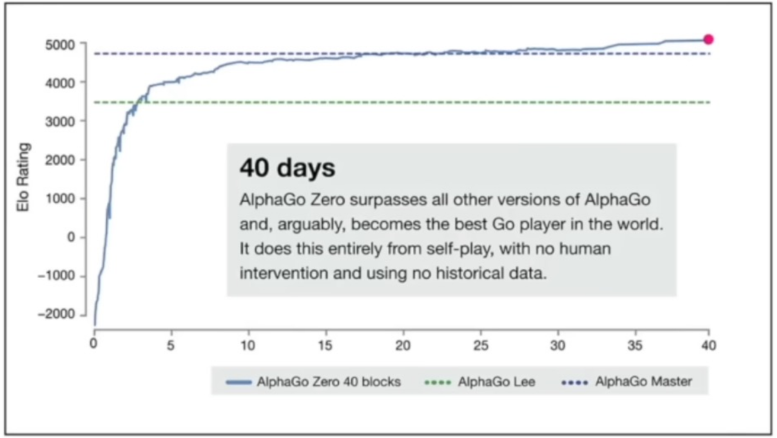
RLHF: REINFORCEMENT LEARNING FROM HUMAN FEEDBACK



Why not use Large Language Models instead

[Andrej's talk on language models]

SELF-IMPROVEMENT (DISTILLATION)



USING THE KNOWLEDGE OF THE MODEL TO IMPROVE THE MODEL

[Andrej's talk on language models]

THANK YOU!