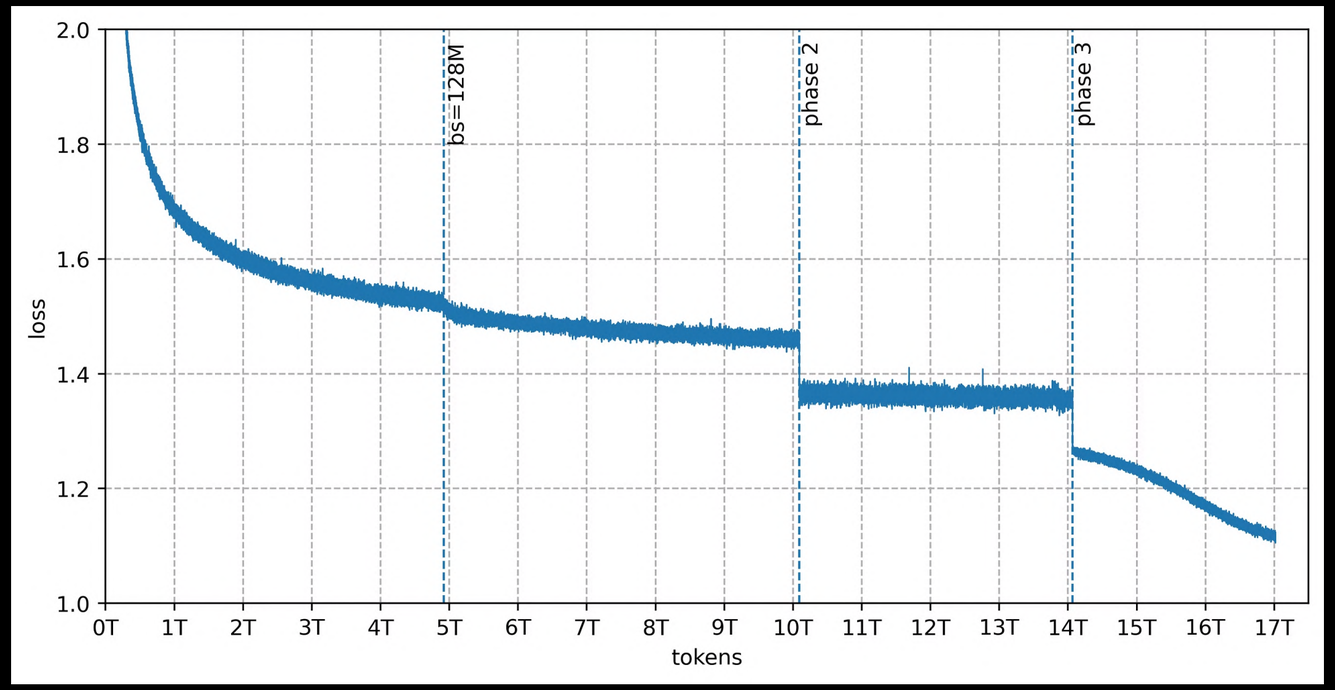
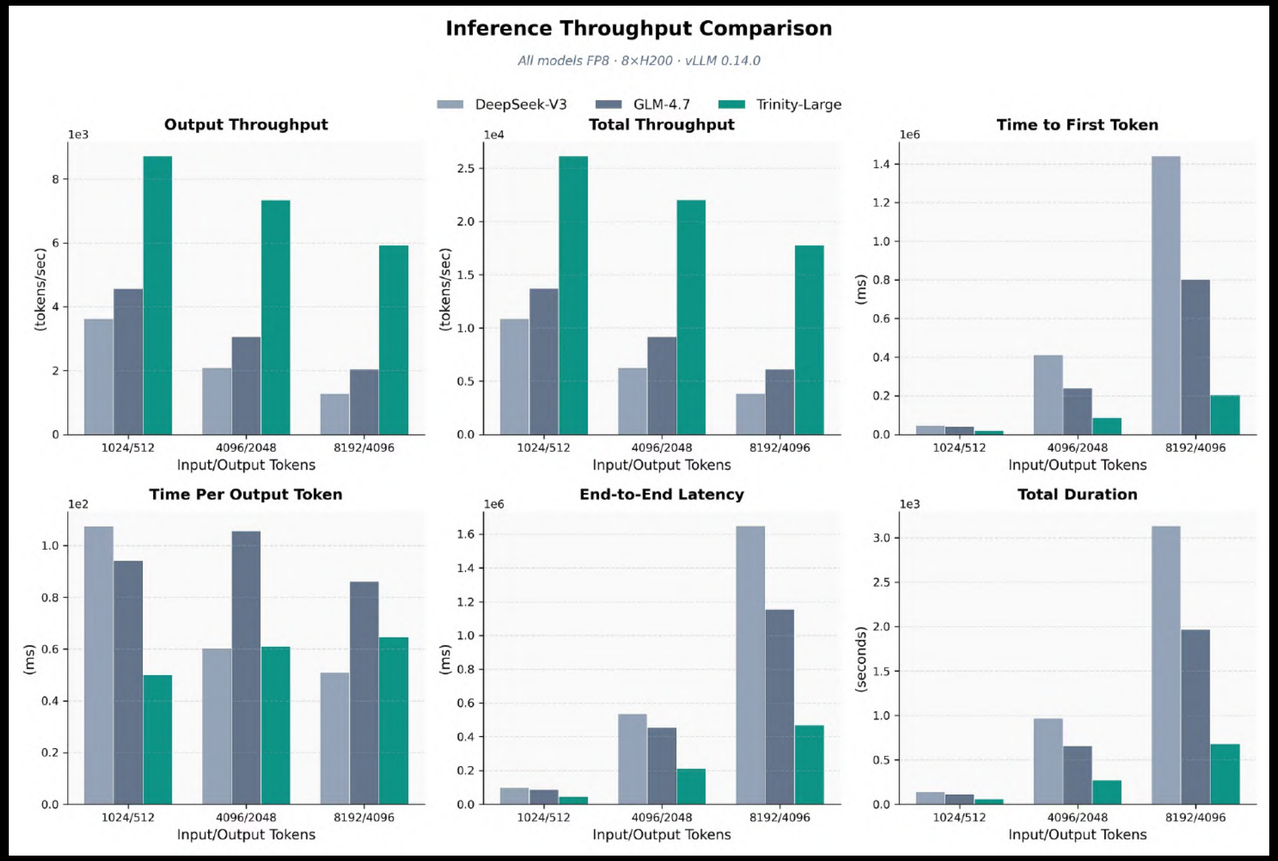

datology ai
Arcee AI
PRIME Intellect
Trinity Large & torchtitan
Matej Sirovatka

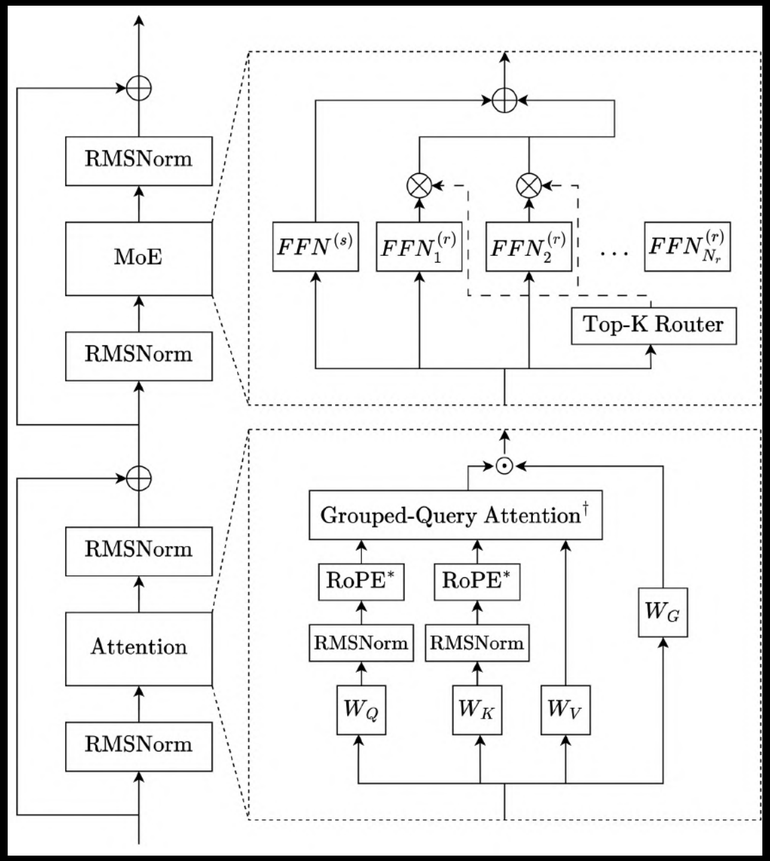
TRINITY LARGE

- 400B/13B
- 4/256 experts
- 1 month
- 2048 B300s
- 20T tokens
- torchtitan only



INFERENCE FRIENDLY





WHAT PROBLEMS CAN WE POSSIBLY HAVE

- problem 1 - 2k B300s
- problem 2 - 3072 hidden dim
- problem 3 - 4/256 experts

HOW TO PARALLELIZE

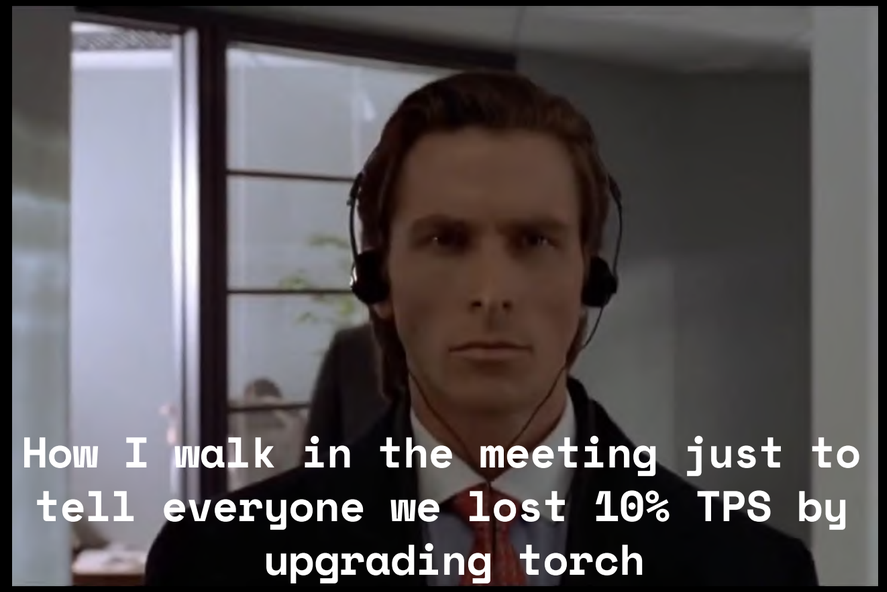
Federico Cassano
@ellev3n11

tensor parallelism is a scam

“Was it wise to pipeline? As we now know, pipelining is not wise.”

SETTING UP ENV IS ACTUALLY HARD

- you upgrade torch and lose 10% MFU
- how to actually enable FA?



FAULT TOLERANCE

- being first to use a cluster is not so nice in general
- torchFT - nice in theory
- heartbeat monitoring & cold-starts

NOW WE CAN JUST TRAIN THE MODEL, RIGHT?

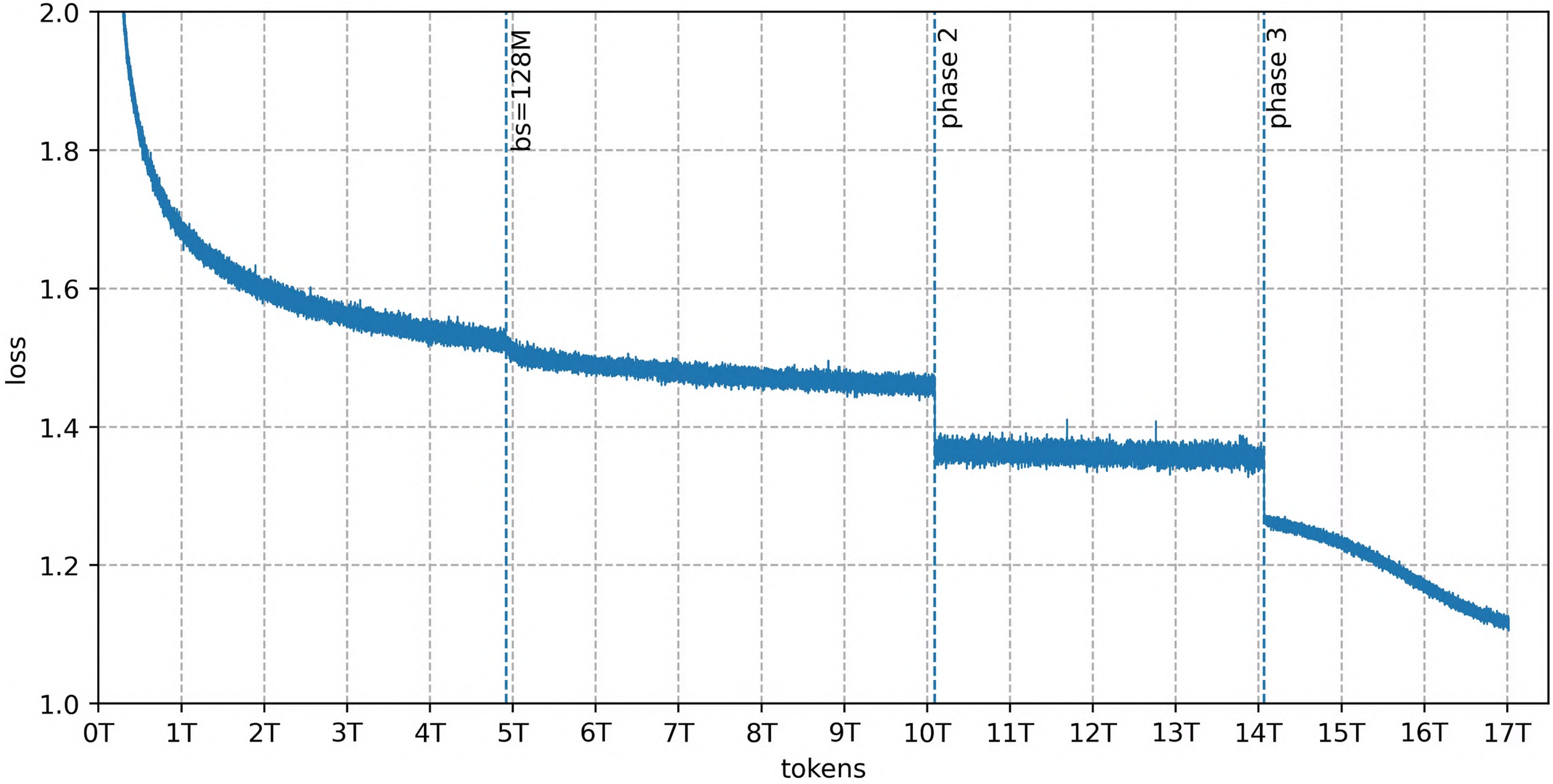
- random expert collapse
- random grad norm spikes

Solution?

- one golden goose run

CONCLUSION?

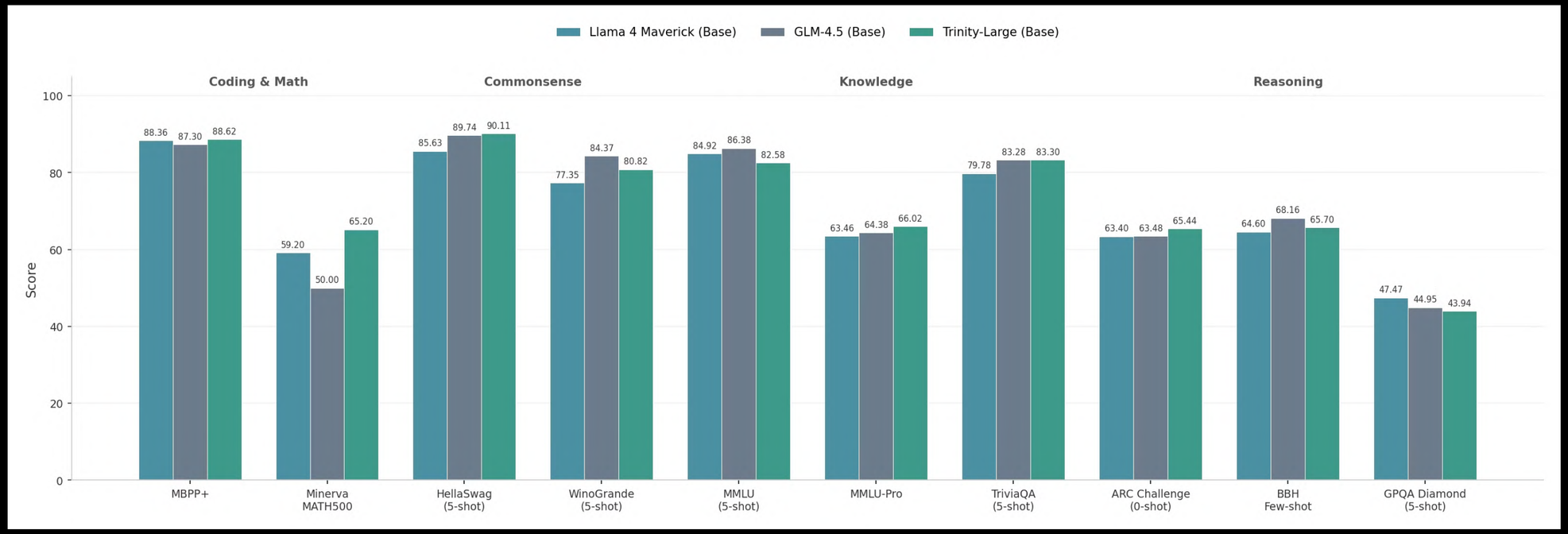
- we managed to train a very sparse model, on a not so stable cluster, with limited time and resources, with a codebase we didn't write
- and it was actually very pleasant


2.0
1.8
1.6
1.4
1.2
1.0
loss
bs=128M
phase 2
phase 3
0T
1T
2T
3T
4T
5T
6T
7T
8T
9T
10T
11T
12T
13T
14T
15T
16T
17T
tokens

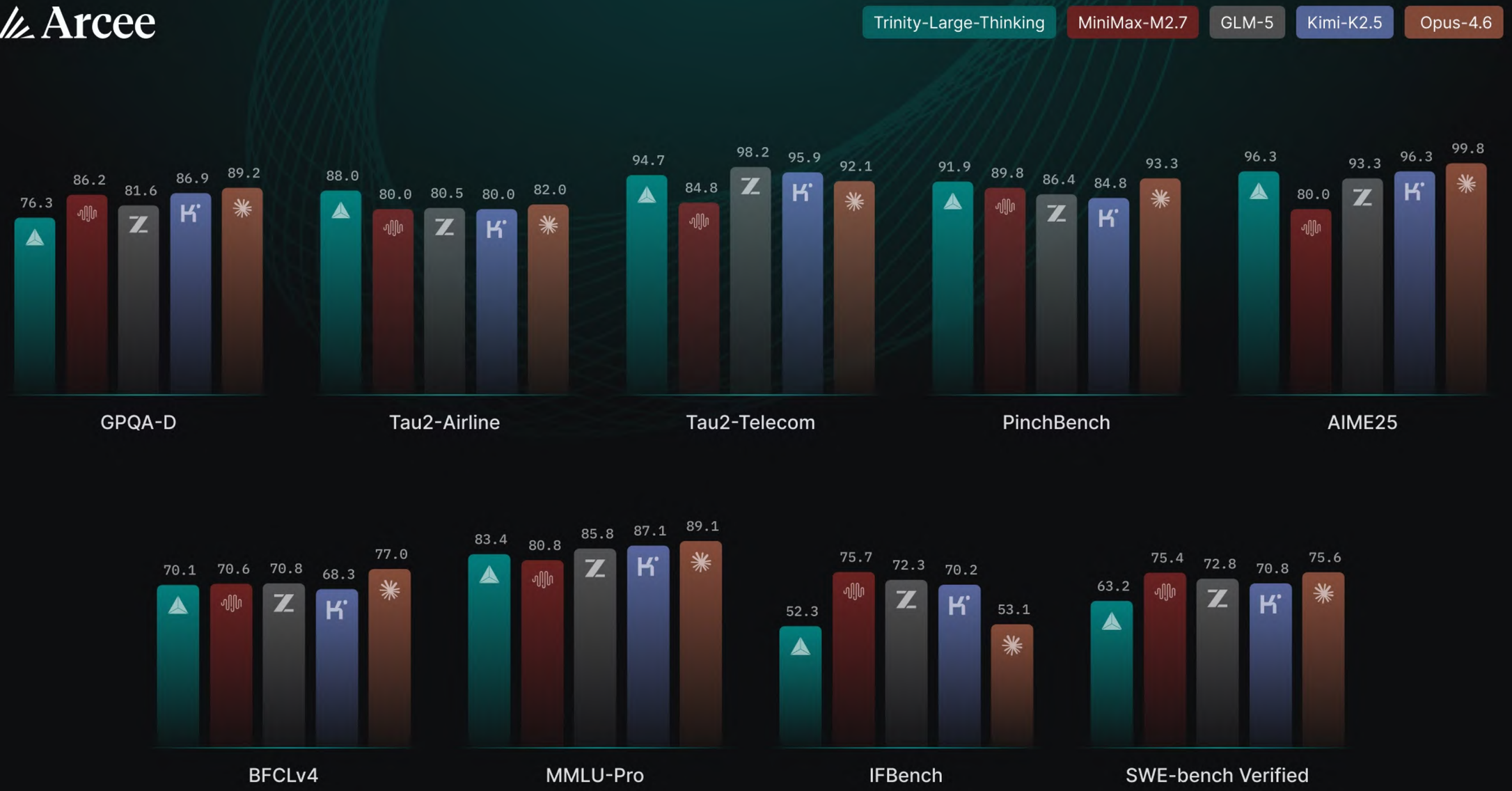
3T tokens on open-router in 50 days

top 4 model on openclaw usage

successful SOTA post-train done by Arcee



Arcee


Trinity-Large-Thinking
MiniMax-M2.7
GLM-5
Kimi-K2.5
Opus-4.6
76.3
86.2
81.6
86.9
89.2
GPQA-D
88.0
80.0
80.5
80.0
82.0
Tau2-Airline
94.7
84.8
98.2
95.9
92.1
Tau2-Telecom
91.9
89.8
86.4
84.8
93.3
PinchBench
96.3
80.0
93.3
96.3
99.8
AIME25
70.1
70.6
70.8
68.3
77.0
BFCLv4
83.4
80.8
85.8
87.1
89.1
MMLU-Pro
52.3
75.7
72.3
70.2
53.1
IFBench
63.2
75.4
72.8
70.8
75.6
SWE-bench Verified