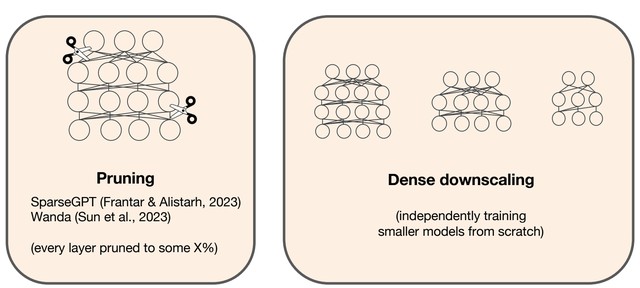
The Cost of Down-Scaling Language Models:
Fact Recall Deteriorates before In-Context Learning

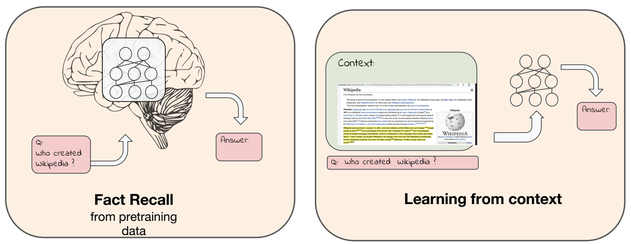
Tian Jin* (MIT), Nolan Clement* (MIT), Xin Dong* (Harvard),
Vaishnavh Nagarajan (Google Research), Michael Carbin (MIT), Jonathan Ragan-Kelley (MIT),
Gintare Karolina Dziugaite (DeepMind)

How does scaling affect LLM capabilities?

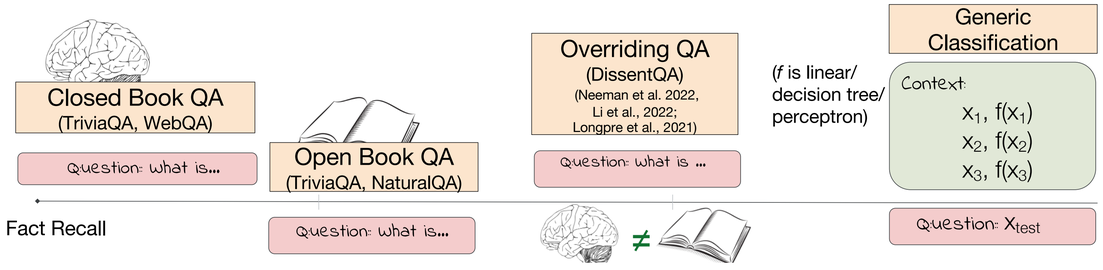
How does scaling number of parameters affect core capabilities of LLMs? We focus on two natural scaling techniques: **pruning** and **dense down-scaling**.



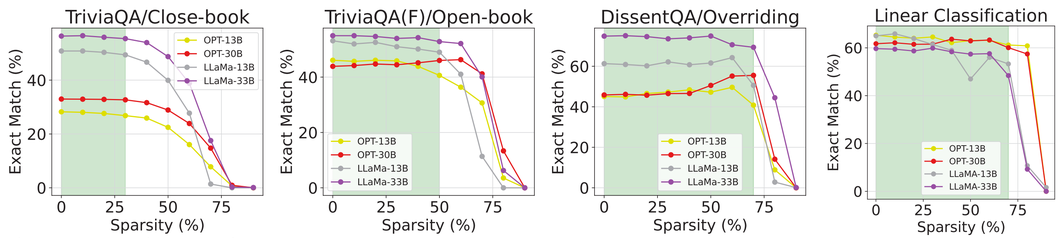
We investigate their effects on two complementary capabilities of LLMs: **fact recall** and **in-context learning.**



But first, how to disentangle fact recall from ICL?



Effect of Pruning

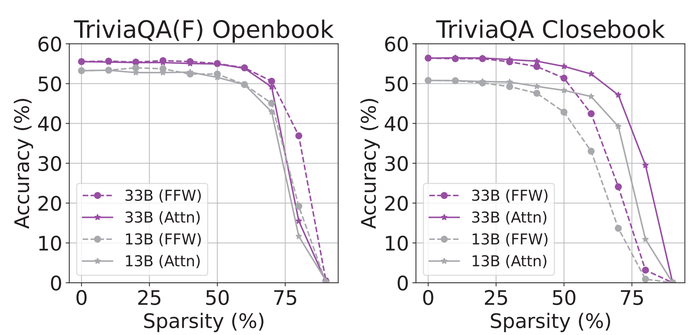


Fact recall deteriorates quickly (5% drop around 30% of pruning).

ICL withstands 60-70% pruning!

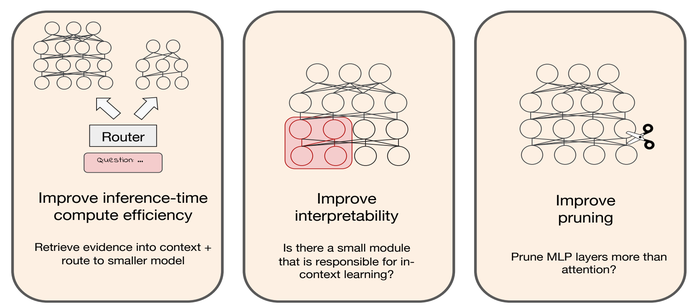
Effect of Pruning Attention vs. FFW

We prune LLaMA 13 and 33B models to find out.



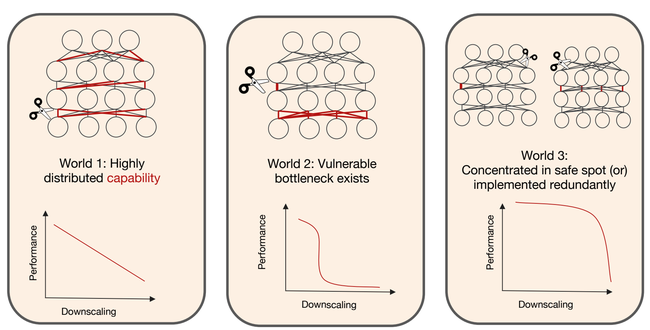
For fact recall, MLP layers are more important than Attention. For in-context learning, not as much difference.

Many implications!

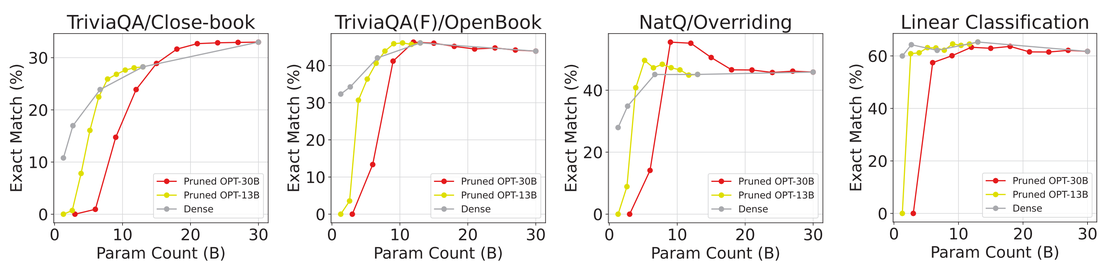


What *could* happen?

There are many possibilities, take your pick now!



Effect of Dense Scaling



Even under dense downscaling, fact recall deteriorates much quicker than ICL, like pruning.

Find our paper on Arxiv!