Overtrained Language Models Are Harder to Fine-Tune

Understanding Scaling Laws in Language Model Training

PRESENTED BY

Abdullah

Introduction and Motivation

Paper: Overtrained Language Models Are Harder to Fine-Tune

Does scaling pre-training always improve model performance?

For years, the dominant belief in AI research has been:

- More data → longer training → better models
- Supported by widely cited scaling laws
- And validated repeatedly in large-model benchmarks

But large language models are rarely used in their raw form.

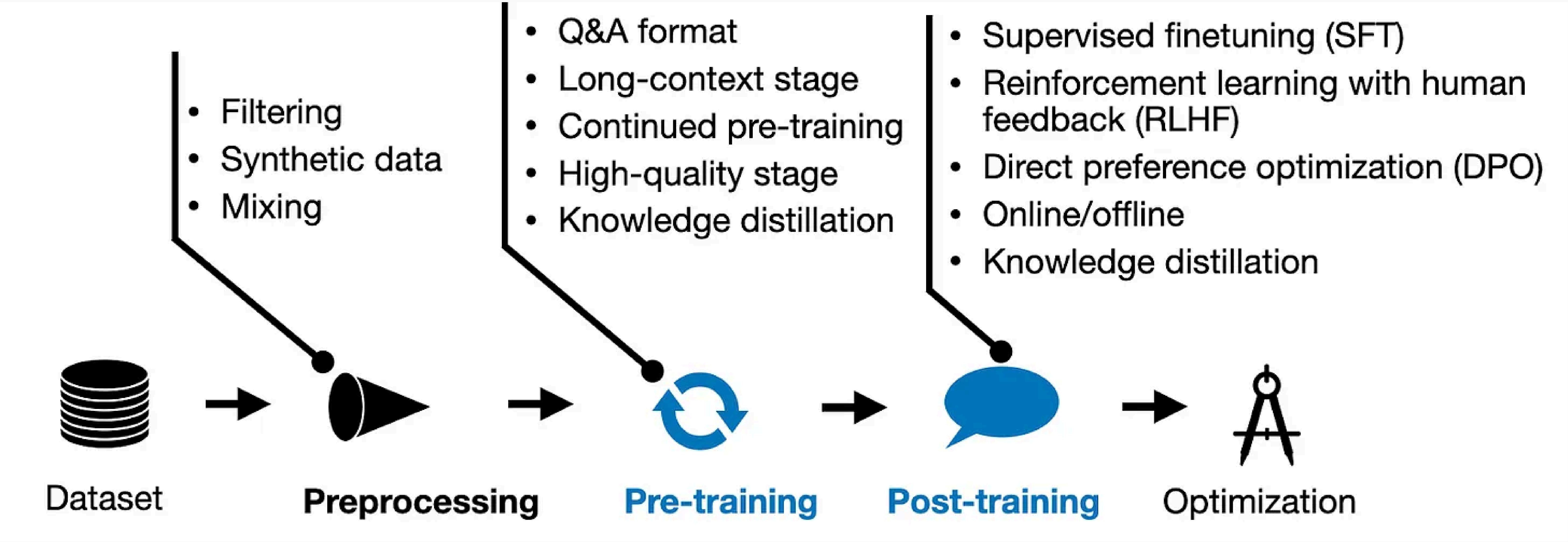
They are almost always:

- Fine-tuned (instruction-following, safety)
- Adapted (domain-specific tasks)
- Aligned (RLHF, SFT, DPO, etc.)

This raises a critical overlooked question:
Does maximizing pre-training performance always lead to better post-fine-tuning performance?

Pre Training vs Post Training

Pre-training builds a model's broad, foundational knowledge from massive datasets (like the internet), teaching general language understanding, while post-training refines that general model for specific tasks, safety, and helpfulness using targeted data and techniques like fine-tuning or Reinforcement Learning from Human Feedback (RLHF), transforming it from a generalist into a useful, aligned tool



SOURCE: HTTPS://MAGAZINE.SEBASTIANRASCHKA.COM/P/NEW-LLM-PRE-TRAINING-AND-POST-TRAINING

Why Revisit Scaling Assumptions?

Three major motivations drive this research:

Real-world models are always fine-tuned:

- Even state-of-the-art models (GPT, Gemini, Claude, Llama) deploy after:
  - Instruction tuning
  - Safety alignment
  - Domain-specific adaptations

So raw pre-training performance is not the final goal.

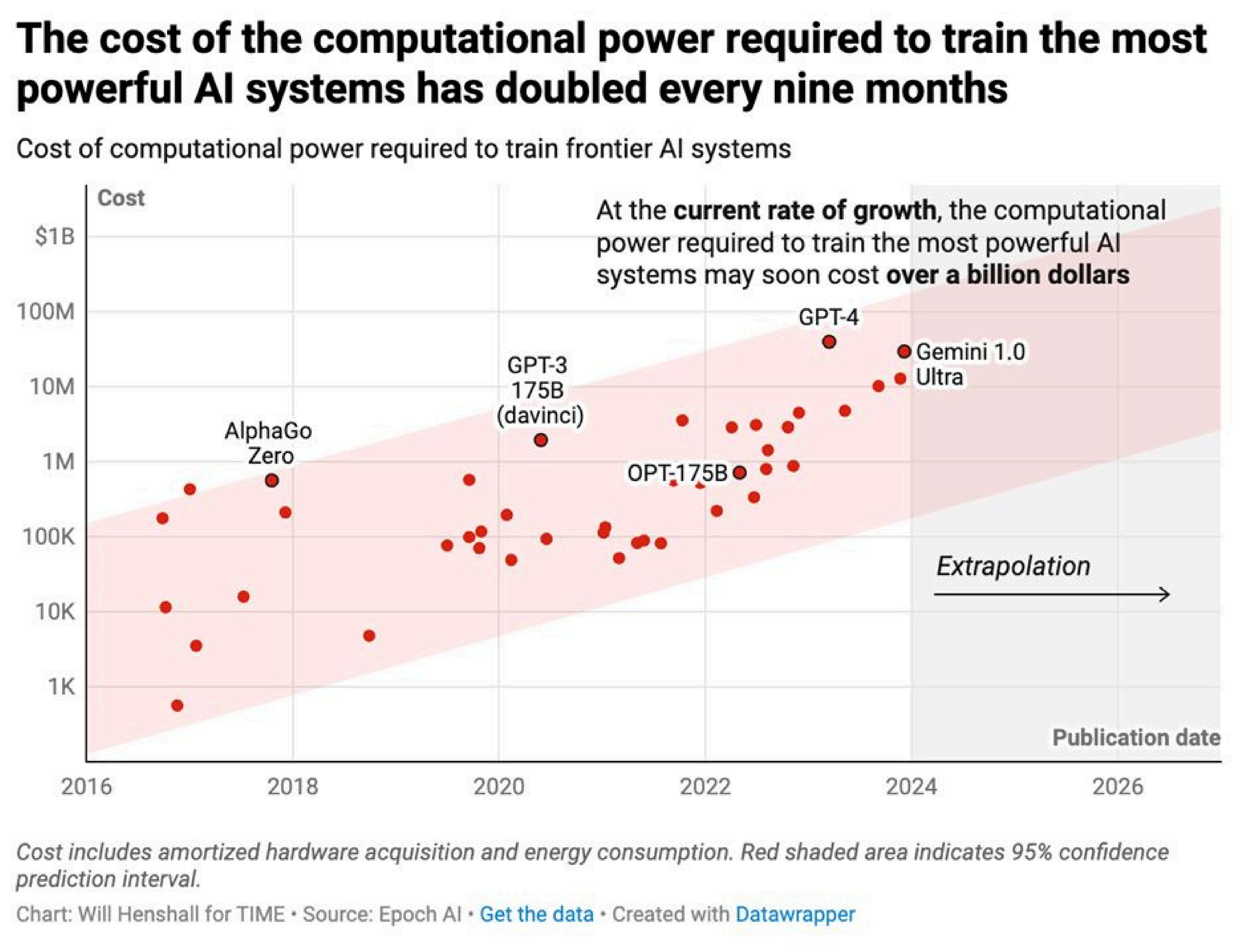
Increasing pre-training cost is enormous

- Training from 2T → 3T tokens increases compute dramatically
- But companies assume this investment is worthwhile
- The paper questions whether it actually helps downstream utility.

Suspicious empirical patterns

- Researchers observed:
  - Longer-trained models sometimes underperform after fine-tuning
  - Even though their base model scores improved
  - This contradicts the standard scaling intuition.

Why Revisit Scaling Assumptions?



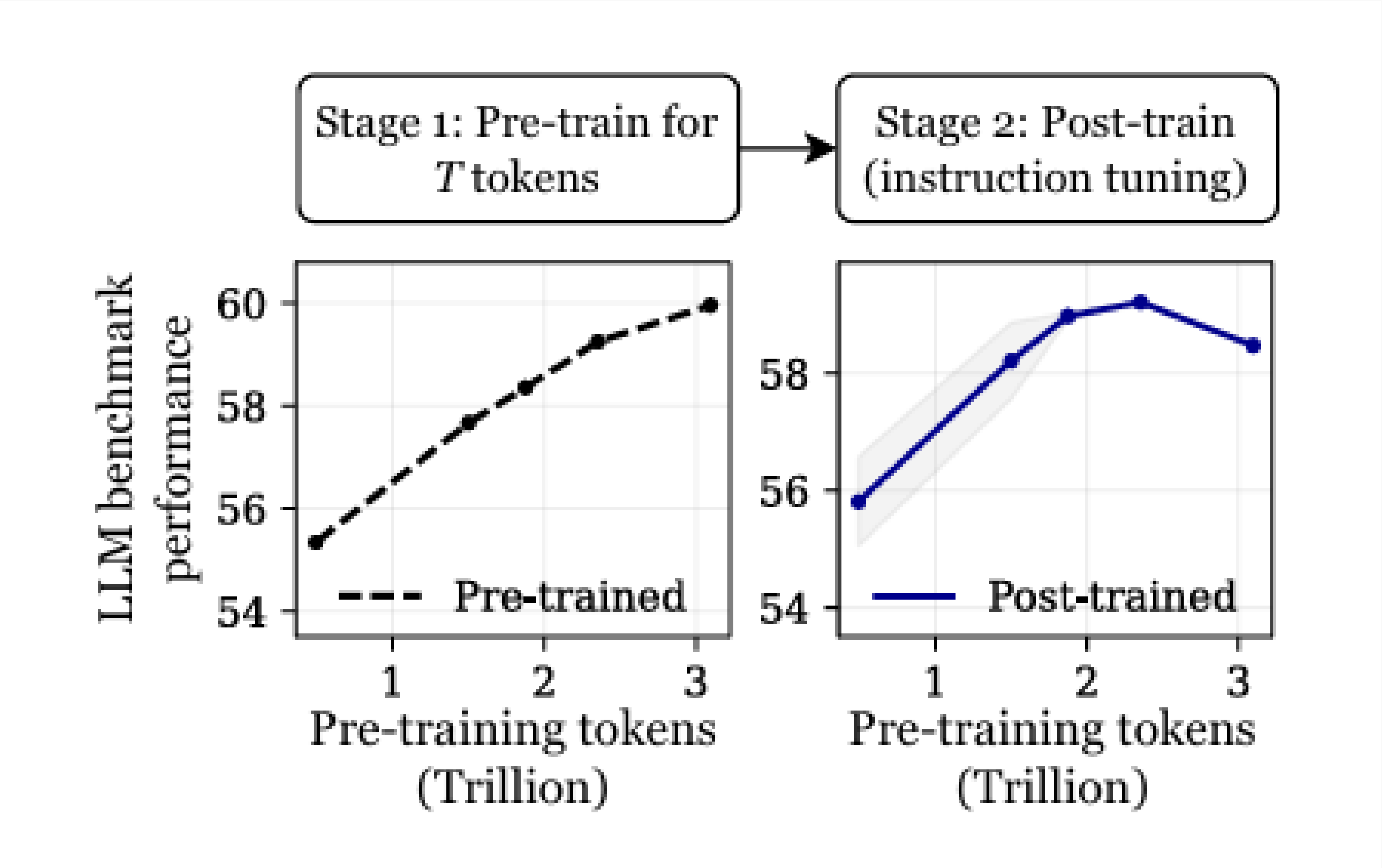
Research Question

Primary question explored in this paper:

How does increasing pre-training length affect performance after fine-tuning?

Sub-questions:

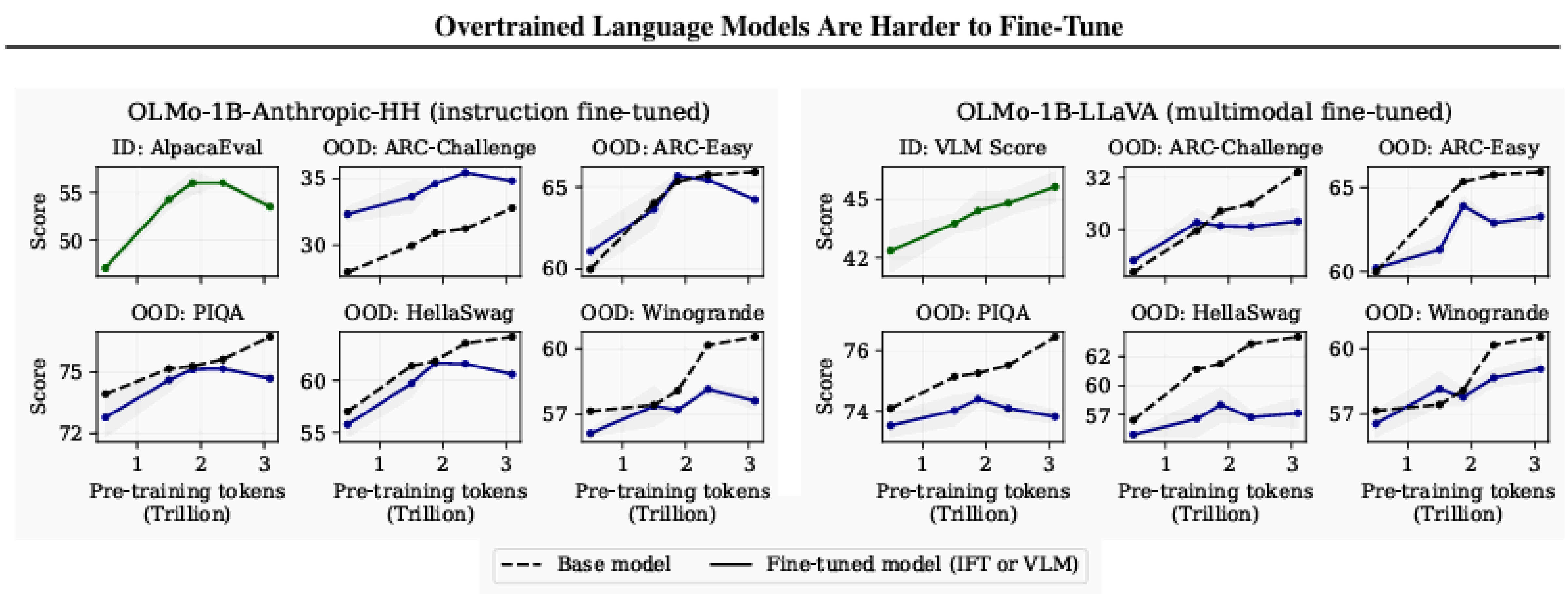
- Does longer pre-training always improve fine-tuned performance?
- Are some models more sensitive to fine-tuning than others?
- What mechanisms cause degradation after fine-tuning?



Observations

More pre-training improves base performance…
…but then reduces performance after fine-tuning.

Example:
As pre-training is extended, the base OLMo-1B model keeps improving, but its fine-tuned performance drops on both instruction-tuning (Anthropic-HH) and multimodal (LLaVA) tasks. This degradation appears across in-distribution and out-of-distribution benchmarks (AlpacaEval, ARC, PIQA, HellaSwag, Winogrande). Results averaged over three runs show that more pre-training consistently makes fine-tuning worse, a trend also confirmed in larger models and additional datasets.



Why does this happen?

The authors propose a mechanism called:

Progressive Sensitivity

As pre-training continues:

1. The model becomes more specialized
2. Internal representations sharpen
3. Parameter space becomes more sensitive
4. Small fine-tuning updates cause large distortions

In short:

Longer pre-training increases sensitivity to parameter updates

Why does Sensitivity Increase?

- Section 4 analyzes a two-layer linear model

- Training learns features incrementally:
    - Strong, robust features first
    - Weak, fragile features later

- Late-learned features:
    - Have small singular values
    - Are easily disrupted by parameter updates

Overtraining = learning features that are increasingly easy to disrupt

The Fine-Tuning Effect

- Fine-tuning modifies model's parameters

- In early-trained models:
  - Updates stay within stable regions

- In overtrained models:
  - Updates overwrite fragile, late-learned features

- Result:
  - Pre-training representations are distorted

Fine-tuning becomes destructive, not because it fails, but because the model is brittle..

Catastrophic overtraining: What is it?

- Catastrophic overtraining occurs when:
  - Sensitivity growth outweighs pre-training improvements
  - Fine-tuning causes severe performance degradation

- Key characteristics:
  - Base model performance keeps improving
  - Post-fine-tuning performance degrades

- This is not catastrophic forgetting
  - The issue is loss of robustness, not loss of capacity

More pre-training makes the final model worse after fine-tuning....

Catastrophic overtraining: Inevitable?

- Theory shows:
  - Sensitivity grows unavoidably with extended pre-training
- If pre-training continues without constraints:
  - Catastrophic overtraining is inevitable
- But with regularization:
  - The inflection point can be delayed
  - But downstream adaptation is reduced
- This is a structural limitation, not a training issue

There is an unavoidable trade-off between robustness and adaptability

Implications

- Scaling laws are incomplete
  - Pre-training tokens do not monotonically improve downstream performance

- Pre-training budgets should be optimized, not maximized
  - More isn’t automatically better

- Fine-tuning pipelines may need redesign
  - Particularly for large, deeply-trained models

- Early stopping during pre-training may improve final usability
  - Strategic choice, not a failure

Limitations

- Evaluated only OLMo architectures
- Limited set of fine-tuning tasks
- Pre-training objective fixed (next-token prediction)
- Results may vary with:
  - Different optimizers
  - Different architectures
  - Alternate pre-training objectives

“Scaling laws don’t guarantee fine-tuning success”

More pre-training improves general capabilities, but can harm fine-tuning performance due to progressive sensitivity.

We must rethink how much data we train, and how we adapt models afterwards.

Thank You