Inference-Time Compute

inference-time-compute · compute · concept

Source: https://policywindow.org/wiki/inference-time-compute

Generated 2026-07-09T20:34:34 UTC

Summary

The scaling regime in which model capability is increased by spending more compute at inference time (multiple samples, search, longer reasoning chains, tool-using iteration) rather than by training a larger model — disrupting the training-compute-as-capability-proxy assumption underlying most current AI governance.

At a glance

Used by

1 instrument(s)

Related concepts

compute-threshold, frontier-tier, capability-elicitation, model-distillation-risk, agentic-system

Primary source

Snell, C., Lee, J., Xu, K., Kumar, A. (2024), 'Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters' — establishes inference-time-compute scaling as a first-class capability lever.

Details

The dominant assumption underlying compute-threshold regulation (EU AIA Art. 51, US EO 14110 §4.2(a)) is that training compute correlates with deployment capability. Inference-time-compute scaling complicates this: a model trained at compute level C can be deployed with inference-time compute K·C per response, producing capability properties intermediate between the base model and a model trained at K·C. OpenAI's o1 (Sep 2024) and o3 (Dec 2024) series, Anthropic's extended-thinking modes, DeepMind's AlphaCode-2 / AlphaProof, and DeepSeek-R1 (Jan 2025) demonstrate the regime empirically. Snell et al. (2024, 'Scaling LLM Test-Time Compute Optimally') and Brown et al. (2024) provide the empirical scaling laws.

Governance implications are direct. (a) Compute thresholds based on training-FLOPs alone (EU AIA 10^2u, US EO 10^2v) understate the deployed capability of inference-scaled models. (b) DeepSeek-R1 demonstrated frontier-tier reasoning at training-compute well below 10^2u FLOPs, weakening the threshold's empirical defensibility. (c) Capability evaluations must specify the inference-compute budget under which the model was tested, since a model can be safe at K=1 and dangerous at K=100. (d) The mitigation surface for inference-time-scaled capabilities is different — restricting access to high-compute deployment APIs is policy-tractable in a way that restricting model-weight distribution is not. The Seoul Declaration + Frontier AI Safety Commitments (May 2024) gesture at this with 'pre-deployment evaluation under realistic conditions,' but no regulator has yet formalised inference-compute-aware thresholds.

How to cite this article

APA

Policy Window. (n.d.). Inference-Time Compute [Wiki article — Concept]. https://policywindow.org/wiki/inference-time-compute

CHICAGO

Policy Window. n.d.. "Inference-Time Compute." Wiki article (Concept). https://policywindow.org/wiki/inference-time-compute.

HARVARD

Policy Window (n.d.) 'Inference-Time Compute', Wiki article — Concept, available at: https://policywindow.org/wiki/inference-time-compute.

OSCOLA

Policy Window, 'Inference-Time Compute' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/inference-time-compute> accessed [date].

BIBTEX

```
@misc{policywindow-inference-time-compute,
title  = {Inference-Time Compute},
author = {Policy Window},
year   = {n.d.},
howpublished = {inference-time-compute — compute},
url    = {https://policywindow.org/wiki/inference-time-compute},
note   = {Primary source: https://arxiv.org/abs/2408.03314}
}
```