ANDYGROVE.SYSTEMS CONTINUING PROFESSIONAL UPSKILLING BRIEF // PP-11 (arXiv:2408.03314)
Open-Access Executive & Engineering Study Guide on Fresh AI Operations Preprints
Full Preprints Library: https://andygrove.systems/preprints
42-Hour CPD Curriculum: https://andygrove.systems/education
Canonical Brief URL: https://andygrove.systems/preprint/pp-11
Primary Source URL: https://arxiv.org/abs/2408.03314

TITLE: Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
AUTHORS: Charlie Snell, Jaehoon Lee, Kelvin Xu, Aviral Kumar (Aug 2024 * UC Berkeley / DeepMind)
PILLAR: TEST-TIME COMPUTE ALLOCATION
---
1. EMPIRICAL & MATHEMATICAL FINDING FOR CONTINUING PROFESSIONAL UPSKILLING:
Shows that allocating inference compute adaptively by prompt difficulty via process verifiers lets smaller models beat 14x larger models at substantially lower total FLOPs.
---
2. ACTIONABLE COO & INDUSTRIAL ENGINEERING PRODUCTION RULE:
Allocate extra verification tokens strictly to hard tail tasks rather than paying frontier parameter prices on 80% routine volume.
---
3. SELF-STUDY & TEAM UPSKILLING INTEGRATION (GOOGLE CLASSROOM & COLAB):
* Interactive Calculators (L0-L3 Wilson TRM & L = lambda*W): https://andygrove.systems/calculators
* Runnable Python Queueing & Routing Colab Lab: https://andygrove.systems/colab.ipynb
* 1-Page Agent Line Audit Specimen: https://andygrove.systems/audit
---
NODE 01 (Part-Time Remote IE/OR/BizOps Pods, $20-45/hr): https://andygrove.systems/#apply
NODE 02 (COO Office Audit Desk): mailto:output@andygrove.systems
Independent platform; not affiliated with Intel Corporation or the Grove family.