ANDYGROVE.SYSTEMS CONTINUING PROFESSIONAL UPSKILLING BRIEF // PP-15 (arXiv:2305.05176)
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-15
Primary Source URL: https://arxiv.org/abs/2305.05176

TITLE: FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance
AUTHORS: Lingjiao Chen, Matei Zaharia, James Zou (Stanford University) (May 2023 * Stanford (TMLR))
PILLAR: LLM CASCADES & CACHING
---
1. EMPIRICAL & MATHEMATICAL FINDING FOR CONTINUING PROFESSIONAL UPSKILLING:
Introduces the foundational three-lever inference cost stack - prompt compaction, semantic/prefix approximation caching, and learned LLM cascades - matching GPT-4 accuracy with up to 98% cost reduction.
---
2. ACTIONABLE COO & INDUSTRIAL ENGINEERING PRODUCTION RULE:
Stack immutable prefix caching (EQ-10) ahead of learned confidence cascades so routine volume never hits frontier billing.
---
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.