CURE-Bench NeurIPS 2025

Dynamic Weighting Improves Model Response Accuracy

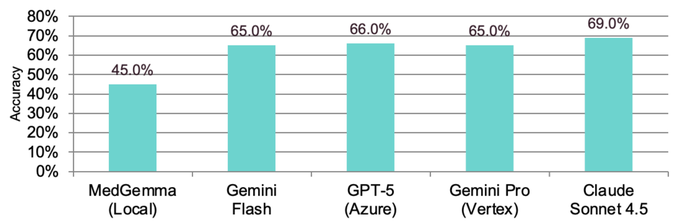
Vivek Kohar, Sahil Bodke UniBio Intelligence

CUREBench Track 1: The Model Gauntlet

- Started with small LM (MedGemma), transitioned to commercial models. Small models (quantized) made simple reasoning errors.
- Low consistency among models and model responses for similar questions

Consistency among models

60-70%



CUREBench Track 2: The Tool Universe

Having knowledge isn't enough

Models need to know WHEN and HOW to use external tools

Tools

~14 MCP servers
OpenFDA, ClinicalTrials.gov, Open Targets, etc.

- **Prompt optimization is crucial for small LM models.**
  - Tool calling efficiency increased three-fold for Nemotron-9B
  - Almost no tool usage by SeedOSS – 36B
  - Gemini Flash-lite excelled in tool calling and was used for tool filtering
  - RAG improved response consistency.
- **Tool calls are necessary but expensive**
  - Prioritized tool calling for inconsistent answers. Google search grounding improved results.
  - Quite a few cases where all models concurred on wrong answers without tool calls.

Discovered highly skewed distribution

70%

of tool calls came from ONE tool

Fewer than 15 tools were used

High disagreement → GPT-5

Ensemble Methods improved accuracy

- Simple majority voting among three models increased accuracy to 72%.
- Implemented a dynamic weighted voting – pick response if all three models agree, continue sampling responses until responses exceed a threshold.
- Tuned weights for different models.

Static Voting

When all models agree → use it

When they disagree → majority wins

Better than any single model ✓

Dynamic Weighted

Claude Sonnet 4.5: Higher weight

Gemini: Lower weight

Additional sampling on disagreements

What Didn't Work: AI Debate Club

Hypothesis: Multi-Agent Discussions

"If we let models debate each other, they'll converge on better answers!"

Reality: Peer Pressure Gone Wrong

- Models gave excessive weight to others' answers
- Frequently changed CORRECT answers to WRONG ones
- Using Claude Opus as "judge" didn't help

https://unibiointelligence.com