

# UniBio Intelligence's Dynamic Weighting Approach

## CUREBench@NeurIPS 2025

---

Vivek Kohar, Sahil Bodke

<https://unibiointelligence.com>

# UniBio Intelligence: Building AI-Native Infrastructure for Biologics Discovery

## UbiMCP

Private Preview

Connect any LLM to biologics databases and tools via MCP servers

## UbiTools

Private Preview

Interactive web platform for state-of-the-art biologics workflows

## UbiChat

Private Preview

Enterprise chat with pre-configured tools + expert access

### Public Database Access

Query UniProt, PDB, and other biologics databases directly from your chat interface.

### Computational Tools

Run sequence analysis, structure prediction, and antibody modeling without leaving your conversation.

### Your Workflow, Your Client

Works with any MCP-compatible chat application. Keep using the tools you love.

- streamlined access to >15 public databases like openfda, opentargets etc.
- ~10 tools
- Search, plan, execute architecture

The screenshot shows the UBI Tools web platform. At the top, there are navigation links for 'UBI Tools', 'Projects', 'Tools', and 'Workflows'. The main heading is 'UBI Tools' with a subtitle 'Interactive biologics design and analysis platform powered by cutting-edge AI tools'. Below this, there are several tool cards: BoltzGen (AI-powered protein binder design tool), ImmuneBuilder (Antibody structure prediction), AbNumber (Antibody numbering & humanization), dyMEAN (Full-atom antibody design), and Sequence Engineering (Antibody sequence analysis & optimization). At the bottom, there are workflow cards for Sequence Annotation and Humanization.

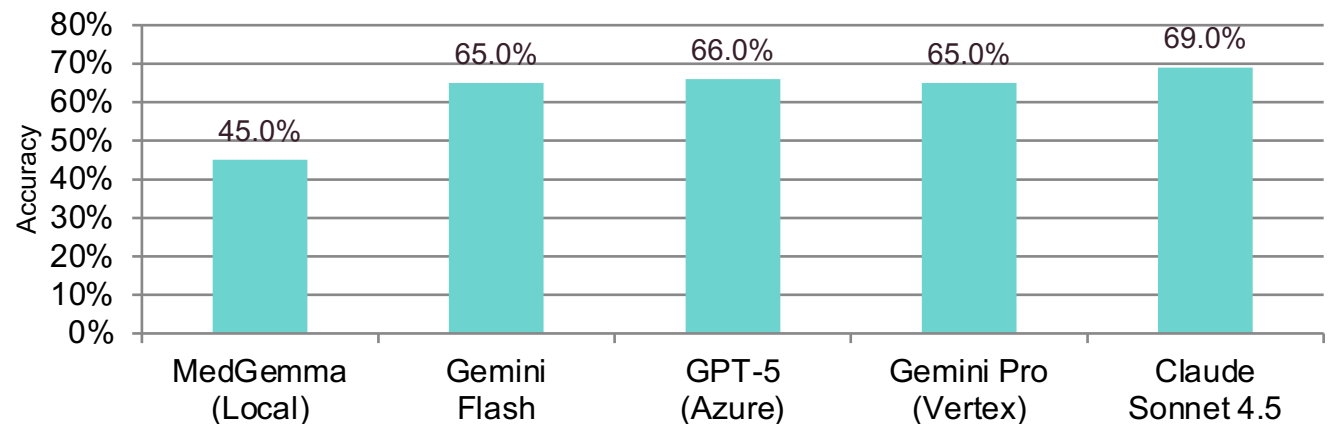
The screenshot shows the UbiChat chat interface. The top bar includes 'Home', 'Chat', 'Workspace', 'Admin Dashboard', and 'Chat History'. The main chat area displays a welcome message: 'Welcome, Vivek Kohar'. Below the message, there is a search bar and a 'RECENT' section with 'No chats yet'. A 'Suggested' section shows 'Antibody CDRs' with a sub-item 'cdrs and numbering schemes'. At the bottom, there are buttons for 'Tools UI', 'Documentation', and 'Contact Support'. The user's name 'Vivek Kohar' and an 'Upgrade' button are visible in the bottom right corner.

# 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%**



# 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

# 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

# Technical Challenges: The Boring But Important Stuff

---

- LiteLLM token tracking inconsistent across providers – additional issues with reasoning traces.
- Safety filters often flag medical content.
- JSONL file format much more convenient than csvs.
- Claude Code turned out to be a pleasant surprise for data analysis.

# Thank you!

---

## Learnings:

- Ensemble methods significantly outperform single models.
- Data-driven tool and prompt selection (GEPA?) beats intuition