# topAI

# A New Class of AI Architecture

M. Tony Ewing 2025-05-07
DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661

# The Problem

## Existing AI Architectures Are Fragile

**Transformers like GPT and DeepSeek's MoE suffer from:**

- **Recursive instability:** *Feedback loops degrade model reliability*
- **Cost Explosion:** *Retraining and scaling require exponential compute*
- **Synthetic data contamination:** *Degraded learning from model-generated inputs*
- **Opacity and drift:** *Unpredictable behavior under distributional shift*
- **Critical use unreliability:** *Mislead decision-makers in high-risk domains—e.g., physicians in differential analysis, autonomous operators under stress, military personnel during target recognition, etc.*

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# The Transformer “Tri-Lemma”

## During development and innovation, Transformers face 3 Challenges:

1. Effectiveness (AI Models Compete on Performance benchmarks)
2. Cost (Training + Inference)
3. Stability (Under recursion or low data)

*Commercial competition between AI models distills to these three dimensions*

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# The Solution: TopAI (topAI)

## A New Class of AI Architecture Designed For:

- **Structural stability** under recursive use and system feedback
- **Low-cost, non-destructive updating**—no retraining required
- **Modular scaling** that allows integration with transformer backbones or as a standalone alternative
- Addressing the **Tri-Lemma**

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# Architecture: topAI

| Feature | TopAI | GPT/DeepSeek |
|---|---|---|
| Update Cost | Minimal | High |
| Drift Risk | None | High |
| Interpretability | Native | Low |
| Compute | CPU / Edge | Multi-GPU |
| Stability | Built-in | Patched-in |

***TopAI updates without retraining by anchoring learning to topological markers within the data landscape.***

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# topAI Placement

***Compatible with existing transformer systems but NOT dependent upon them:***

- ***Operates as an interference engine*** *(behind or beside LLMs)*
- ***Stabilizes output*** *(with transformers through totemic memory and projection)*

*Or,*

- ***Serves as a standalone architecture*** *in domains where drift and hallucination is fatal*

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# Who This Is For

**TopAI is NOT a chatbot, but a recursive inference engine for:**

- **Healthcare**—Real time diagnostics
- **Finance**—Market regime shift resilience
- **Cybersecurity**—drift-aware monitoring
- **Embedded AI**—wearables, drones, medical devices
- **Safety systems**—constrain LLMs, prevent hallucinations

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# Key Differentiators: topAI

- **20 - 40x cheaper to deploy (100x in some environments)**
- **No retraining required under feedback**
- **Zero hallucination by design**
- Interpretable and memory-preserving
- Runs on CPUs and edge devices

*TopAI replaces statistical approximation with topological memory and constraint—preserving structure rather than flattening it*

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# Early Results: topAI versus GPT Models

| Metric | GPT-Class (GPT-4o / DeepSeek) | TopAI |
|---|---|---|
| **Accuracy (diagnostic task)** | 94–97% (with prompt tuning) | 97% (native, no tuning) |
| **Update Time (CPU)** | N/A or requires retrain (~hours) | 0.008 sec |
| **Retraining Cost (per cycle)** | $1–5M (est.) | $0 (no retraining) |
| **Inference FLOPs (per query)** | 1.1e11 – 1.4e12 | ~2e9 (40–500x cheaper) |
| **Hallucination Rate** | ~3–12% depending on prompt | ~0% in constrained domains |
| **Deployment** | GPU-only / Cloud-dependent | CPU-ready / Edge-compatible |
| **Stability Under Feedback** | Degrades or collapses | Stable by design |

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# Early Results: Cardiac Arrest Prediction Task

**Dataset:** *Cardiac Arrest Events Dataset,* sourced from MMIC-IV Clinical Database (PhysioNet)
Filtered 6,200 patient records with time series vitals and events outcomes

## Model Comparison

| Metric | Random Forest (Baseline) | TopAI Prototype |
|---|---|---|
| **Accuracy (AUROC)** | 96.3% | 97.1% |
| **Update Time** | 0.74 sec (CPU, partial retrain) | 0.008 sec (no retrain) |
| **Inference Time per Sample** | 0.11 sec | 0.006 sec |
| **FLOPs per Update** | ~3.4e9 | ~2.6e7 (130x— cheaper) |
| **Stability Under Recursion** | Degrades | Preserved across 10+ cycles |
| **Deployment Hardware** | GPU/Cloud | CPU/Edge-compatible |

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# topAI: AI Tri-lemma Is Broken

| Model | Performance | Cost | Stability | Overall |
|---|---|---|---|---|
| **GPT-4o** | SOTA | Extreme | Collapses under recursion | Powerful but expensive and unstable |
| **DeepSeek** | High | Lower | Degrades under updates | Cheaper GPT, but still brittle |
| **TopAI** | High | Minimal | Stable by design | Cheaper, safer, and enduring |

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

# Key Takeaways

- **TopAI exhibits equal or better prediction accuracy vs. state-of-the-art models**
- **130x faster and cheaper to update**
- **No retraining required**
- **Coherence preserved** even under simulated patient feedback loops
- **Fully testable on commodity laptop hardware**

**DOI-protected content: Zenodo 10.5281/zenodo.15345660 & 15345661**

## CONTACT

Maurice “Tony” Ewing
TonyEwing.com/BlackCopernicus
ewing@tonyewing.com

[DOI-protected papers: Zenodo 10.5281/zenodo.15345660 & 15345661]