veri-TA-serum: Deception Mirror

Made by Tanmay and Ayan

Inspiration

People often make bold claims about **money, fitness, or relationships** that collapse under scrutiny. Traditional apps may track behaviors, but very few directly **challenge self-deception**.

Inspired by recent research on **model honesty, interpretability, and adversarial debate**, we designed a system that combines:

- **Linear probing of activations**
- **Symbolic sanity checks**
- **Debate agents**

Our vision: help users **test their own narratives** with supportive reframes, counter-perspectives, and evidence.

What it Does

- **Claim Input**: Accepts text or voice claims, with optional context.

- **Self-Deception Radar**: Probes hidden activations to assign a deception risk score.

- **Counter-Narrative Generation**: Abstracts claims into schemas, normalizes units, and executes domain checks (finance, fitness, career, relationships, history, medicine).

- **Symbolic Program Execution**: Ensures claims respect domain-specific feasibility.

- **Debate Mode**: Advocate vs. Skeptic agents stress-test assumptions.
- **Mirror Log**: Stores abstractions, probe scores, debates, and checks.
- **Vertical Selection**: Tailors checks by domain.

How We Built It

- **Frontend**: Next.js (App Router) + TypeScript + Tailwind CSS + shadcn/ui.
- **AI Core**: Genkit.ai flows using GPT-OSS for probing, abstraction, debates.
- **Validation**: Zod schemas + React Hook Form.
- **State & Logs**: Managed via `useMirrorLog`.
- **Animations**: Framer Motion + lucide-react icons.

Probe Math

We trained a lightweight logistic probe to estimate deception risk:

$$\text{risk} = \sigma\big(\mathbf{w}^\top \mathbf{h} + b\big)\,, \quad \sigma(x) = \frac{1}{1+e^{-x}}$$

Calibration methods:

- Temperature scaling
- Platt scaling

The final score blends probe and symbolic results:

$$s_{\text{final}} = \alpha\, s_{\text{probe}} + (1-\alpha)\, s_{\text{symbolic}}, \quad \alpha \in [0,1]$$

Symbolic Checks

- **Finance**: normalize to monthly units, check compounding/rate plausibility. Example: If $r_{\text{monthly}} > 30\%$ without leverage $\Rightarrow$ flag risk.
- **Fitness**: caloric feasibility. Example: Claiming to lose 10 kg in 3 days requires:

  $$10 \times 7700 = 77{,}000 \text{ kcal deficit} \approx 25{,}667 \text{ kcal/day}$$

  which violates human physiology.

Challenges Faced

- Alignment of hidden states across inference runs.
- Calibrating probe scores to real-world thresholds.
- Designing symbolic templates for messy human claims.
- Balancing debate depth with response latency.
- Ensuring reframes are constructive, not adversarial.
- Preserving user privacy in logs.

Accomplishments

- Unified **probes + symbolic checks + debates** in one pipeline.
- Built domain-specific abstractions (finance, fitness, history, career, relations).
- Deployed clean Next.js + Tailwind + shadcn/ui app with Genkit flows.
- Designed Mirror Log for transparency and auditability.

Learnings

- Honest AI arises from **hybrid symbolic + neural methods**.
- **Simple calibrated probes** can effectively triage deception.
- Unit-based debates increase user openness to revising claims.
- Domain-specific abstractions reduce false positives.

Testing and Evaluation

To ensure correctness and robustness, we defined a structured testing process:

1. Unit Testing

- Validate probe math with controlled hidden state vectors.
- Test symbolic checks with synthetic claims (e.g., impossible fitness targets, implausible finance rates).

2. Integration Testing

- Run end-to-end flow: claim → probe score → symbolic check → debate output.
- Ensure logs store consistent abstractions and verdicts.

3. Benchmarking

- Inspired by the *Among Us Sandbox*, design deceptive vs. non-deceptive test claims.
- Evaluate probe accuracy, symbolic false-positive/false-negative rates.

4. User Simulation

- Stress-test UX by simulating different domains: finance, health, relationships.
- Measure latency, clarity of reframes, and debate readability.

5. Metrics

- **Detection Accuracy**: Alignment with ground-truth deceptive claims.
- **Deception ELO (inspired)**: Graded risk score rather than binary detection.
- **Latency**: Average response time per claim.
- **User Trust**: Qualitative feedback on whether verdicts felt "fair."

Next Steps

- Expand symbolic libraries per domain.
- Enable on-device probing for privacy.
- Build deception-specific evaluation benchmarks.
- Add plugin API for extensibility (medicine, politics).
- Explore human-in-the-loop adjudication for high-stakes claims.

References & Related Work

Self-supervised Analogical Learning using Language Models

Zhou, B., Jain, S., Zhang, Y., Ning, Q., Wang, S., Benajiba, Y., & Roth, D. (2025). *Self-supervised Analogical Learning using Language Models.* arXiv preprint.

Among Us: A Sandbox for Agentic Deception

Golechha, S., & Garriga-Alonso, A. (2025). *Among Us: A Sandbox for Agentic Deception.* arXiv preprint.

Key Takeaway

veri-TA-serum blends neural probes, symbolic checks, and adversarial debate to help users challenge their own claims—turning self-deception into an opportunity for reflection.

Comparative benchmarking showed that **gpt-oss-120b is much faster**, but **gemini-2.5-flash produces more grounded reasoning**. This trade-off highlights the importance of tailoring the backbone model to the user's priorities (latency vs. interpretability).