The Ghost in the Machine

Precog Recruitment Task 2026 Submission: Key Findings

Samarth Rao

IIIT Hyderabad

Repository: `github.com/rao-samarth/human-or-ai`

Dataset Construction

Class 1: Human-Written Text

- Sourced 100 books from Project Gutenberg (Conan Doyle, Wodehouse, Twain, Shakespeare).
- Cleaned and split into 20 paragraphs per book (100-200 words each).

Class 2: AI Generated

- Generated 500 paragraphs using Gemini 3 Pro API.
- Used "diversity modes" and Temperature 1.0 to prevent similarity.

Class 3: AI Mimicry

- Prompted Gemini 3 Pro to specifically mimic Class 1 authors.

Class 4: Failed Fine-Tuning

- Attempted Unsloth fine-tuning; failed due to overfitting (high LR) and prompt leakage.

The Detectors

1. The Statistician

- **Models:** XGBoost & Random Forest.
- **Perf:** 85% - 93% accuracy.
- **XGBoost:** Exploited **em-dash frequency** (Importance > 0.28).
- **Random Forest:** Forced to rely on **Hapax Legomena**.

2. The Semanticist

- **Arch:** Word2Vec + MLP.
- **Method:** "Bag of Means".
- **Perf:** 97.46% / 95.91%.
- **Key:** Outperformed Statistician; semantics harder to mimic than punctuation.

3. The Transformer

- **Model:** DistilBERT (Fine-tuned).
- **Acc:** $> 99\%$.
- **Checks:**
- LoRA vs No-LoRA (99% vs 33%).
- Learning Curve (Gradual growth).
- Weights SD (0.0019).

Saliency Mapping: Why the Model Fails

Analysis of edge cases with lowest confidence

Scenario	Reason for Error
Human → AI	Author used overly dramatic/descriptive language (e.g., "turmoil", "countless"), which the model associates with creative AI.
AI → Human	AI used specific, grounded details (e.g., "handsome fee") instead of vague fluff.
Mimicry Success	AI successfully used "classic" human vocabulary like "noble" or "civic beautification".

The MATE Framework

MATE

Memetic Algorithm for Text Evolution

Objective: Treat adversarial text generation as a constrained optimization problem.

Optimization Goal

Maximize $P(\text{human}|x)$ while maintaining:

1. Semantic Similarity
2. Fluency (Low Perplexity)

MATE Methodology: Global Search

1. **Saliency-Guided Reduction:**
   - Identify top 20% of tokens contributing to "AI" classification using gradients.
   - Mark only these as mutable.
   - **Impact:** Reduces search space by 90%.
2. **Initial Population:**
   - Generated via Gemini 3 Pro (Temp 1.0) for variation.
3. **Crossover & Mutation:**
   - Select parents by fitness.
   - Create offspring using Gemini 3 Pro API (merging styles, keeping content).

MATE Methodology: Local Search

Simulated Annealing

- One-by-one perturbation of mutable tokens.
- Allows candidates to escape local optima/plateaus where greedy algorithms stall.

Lagrangian Relaxation

- Dynamic penalties applied to perplexity and semantic similarity constraints.
- If constraints are violated, penalties increase, forcing the next generation to prioritize fluency.

MATE Results

- **Evasion Success:** Evolved text from 8×10^{-7} (approx 0%) to **77.7%** Human Probability.
- **Stabilization:** Crossed "Human" category at 12th iteration; stabilized at $\sim 78\%$.

Method	**Time per Generation**
Standard Evolution	$\sim$ 35 mins
MATE (Search Space Reduced)	$\sim$ **6 mins**

Conclusion & Future Work

Personal Insight:

- Making human text sound like AI is easy (synonyms).
- Making AI sound human is hard; the most effective manual strategy was "rambling".

Future Directions:

1. Use MATE-generated adversarial examples to re-train the detector (GAN-style).
2. Modify MATE to detect and destroy watermarks.
3. Properly implement Unsloth fine-tuning.