Precog - Task Summary

Venya Velmurugan

February 2026

1 Task 0: The Library of Babel

1.1 Overview

I constructed a three-class authorship-controlled dataset where style, not topic, distinguishes classes: (1) Human text from Austen and Gaskell novels, (2) AI-generated text in neutral voice, (3) AI-generated text mimicking Austen/Gaskell style. All paragraphs have normalized to 100–200 words (as required) across nine shared thematic anchors.

1.2 Key Process: Iterative Prompt Refinement

Try 1 (Topics–Lenses–Constraints): Structured prompts with explicit topic, lens, and constraints produced highly uniform AI text with reduced lexical diversity, lower punctuation density, and uniform sentence structure. Task 1 analysis revealed trivial separability.

Try 2 (Story Generation + Segmentation): Generated long narratives then segmented into paragraphs. Improved surface diversity but retained global stylistic coherence across segments, with overly smooth POS distributions and limited syntactic variance.

Try 3 (Few-Shot Paragraph Prompting): Independent paragraph generation with minimal structural guidance, using example paragraphs to convey style implicitly. Successfully matched human text in lexical richness, POS ratios, and syntactic depth.

Sampling Parameters: Progressively increased diversity across tries. Try 3 used temperature=1.3–1.4, top-p=0.96–0.98, top-k=200–250, producing closest alignment with human stylometric distributions.

1.3 Main Results

Dataset refinement confirmed through Task 1 statistical analysis: Try 1 trivially separable, Try 2 showed residual regularities, Try 3 achieved non-trivial separability suitable for meaningful detection experiments. Final dataset: 4000+ human paragraphs, 1000+ AI paragraphs per class.

1.4 Key Insight

Prompt design influences stylistic signals more than model architecture itself. Statistical validation essential to detect artifacts invisible during manual inspection.

2 Task 1: The Fingerprint

2.1 Overview

I extracted 10 stylometric features (lexical richness, syntactic complexity, punctuation density, readability, information-theoretic measures) to verify measurable differences between classes. Used Cohen's d as primary metric rather than p-values due to large sample size (4000+ paragraphs).

2.2 Main Results Across Dataset Versions

Try 1: Large effect sizes across most features. AI text showed artificially inflated MSTTR and hapax ratios (prompt forcing vocabulary exploration), higher adjective-to-noun ratios, deeper dependency trees, much lower punctuation density (large effect sizes for semicolons/em-dashes/exclamations), and significantly harder readability (very large effect size). Classes trivially separable.

Try 2: Reduced separability. MSTTR differences negligible, hapax ratios aligned with expectations (higher for human), syntactic complexity showed medium effect sizes. Punctuation density remained strongly discriminative (large effect sizes). AI text became easier to read than human text post-segmentation. Information-theoretic measures showed lower AI entropy and perplexity as expected.

Try 3: Closest alignment with human text. Lexical richness differences negligible, syntactic complexity minimal (dependency depth nearly identical), punctuation density most discriminative remaining signal (semicolons large effect, exclamations medium effect), readability differences negligible, information-theoretic measures converged substantially.

2.3 Key Insight

Effect size analysis (Cohen's d) more meaningful than statistical significance with large samples. Try 3 achieved design goal: mathematically distinct but not trivially separable, validating dataset for downstream tasks.

3 Task 2: The Multi-Tiered Detective

3.1 Overview

I trained three detector tiers across all the 3 dataset versions: Tier A (Random Forest on 8 stylometric features), Tier B (shallow neural network on averaged GloVe embeddings), Tier C (DistilBERT with LoRA fine-tuning). Evaluated under identical 75-25 splits with class-weighted loss to handle imbalance.

3.2 Main Results

Tier A (Random Forest):

- Try 1 & Try 2: ∼97.5% accuracy, high AI-class precision/recall
- Try 3: 93.5% accuracy, AI F1 ≈ 0.83
- Confirms explicit stylometric signals captured well but degrade with prompt refinement

Tier B (Semantic Neural Network):

- Try 1: 97% accuracy, AI F1 ≈ 0.93 (semantic differences detectable)
- Try 2 & Try 3: Substantial drops, AI recall ∼0.66–0.78, accuracy <90%
- Confirms semantic alignment improves across dataset versions; averaged embeddings insufficient when topic-controlled

Tier C (Transformer with LoRA):

- Consistent >99% accuracy across all tries
- AI-class F1 > 0.97 even for Try 3
- Minimal train-test gap; 60-40 split showed test accuracy exceeding training (strong generalization)
- Contextual and discourse-level cues persist despite minimized surface/semantic signals

3.3 Key Insight

Clear hierarchy: Tier A exploits explicit features (degrades with refinement), Tier B relies on semantics (fails under topic control), Tier C remains robust (captures subtle contextual patterns). Try 3 confirmed as non-trivially separable, suitable for interpretability analysis.

4 Task 3: The Smoking Gun

4.1 Overview

I applied interpretability methods to Try 3 Tier C detector (99.1% accuracy): AI-isms frequency analysis (250+ catalog words), SHAP word attribution, Captum Integrated Gradients, deep error analysis (all 11 misclassifications), Class 2 vs Class 3 comparison, aggregate pattern analysis.

4.2 Main Results

4.2.1 AI-isms Frequency Analysis

- Try 1: Human 1.92±1.86, AI 3.54±2.36, Cohen's d=0.76
- Try 2: Human 1.92±1.86, AI 3.71±2.54, Cohen's d=0.80 (strongest)
- Try 3: Human 1.92±1.86, AI 3.18±2.23, Cohen's d=0.61
- All comparisons $p < 10^{-19}$, frequency ratios 1.7–1.9×

4.2.2 Style Mimicry (Class 2 vs Class 3)

- Try 1: Detection 22.2%→16.1% (−6.1pp), AI-isms 3.92→3.16 (−0.76)
- Try 2: Detection 100%→100% (0pp evasion)
- Try 3: Detection 100%→94.4% (−5.6pp), AI-isms 3.63→2.73 (−0.91)

Style mimicry reduces AI-isms by 0.76–0.91 but provides only 0–6.1pp evasion benefit. Contradicts hypothesis of fundamental detectability shift.

4.2.3 SHAP vs Captum Divergence

SHAP top words: "baker's", "wife—a", "bread", "hearth", "first", "wasn't" (contextual/structural markers).

Captum top tokens: "eleanor", "marianne", character names, formal spellings ("colour", "honour"), semicolons.

Critical finding: *Zero overlap* between top SHAP words and AI-isms catalog. Model learns distributional regularities and contextual patterns, not just suspicious vocabulary.

4.2.4 Error Analysis: The AI-isms Paradox

False Positives (4 human→AI):

- Mean AI-isms: 1.25±0.83 (vs 1.95±1.32 for correct human predictions)
- **FPs contain 35% fewer AI-isms than correctly classified human text**
- Triggered by formality: semicolons (4/4 FPs), character names, complex sentences
- Confidence: 57.7% (vs 99.1% for correct predictions, 41.4pp gap)
- Root cause: Model conflates formal 19th-century register with AI's attempt to emulate formality

False Negatives (7 AI→human):

- Mean AI-isms: 2.00±1.20 (vs 3.18±2.23 for correct AI predictions)
- **FNs contain 37% fewer AI-isms than correctly classified AI text**
- One sample achieved 0 AI-isms (perfect evasion)
- Camouflage: Character names ("martha" 3/7 FNs), period vocabulary ("pottery", "mill", "vicar"), conversational markers ("aye", "remarked")
- Confidence: 65.1% (vs 98.1% for correct predictions, 33.0pp gap)
- Root cause: AI text with < 2 AI-isms + authentic period names/vocabulary evades detection

4.2.5 Aggregate Patterns

- **Confidence gap**: All errors 62.4% vs correct predictions 98.6% (36.2pp gap). Only 11/1264 high-confidence (> 95%) predictions wrong. Model recognizes uncertainty.
- **Optimal AI-isms threshold**: 5.5 AI-isms achieves 80.8% accuracy vs 99.1% neural (18.3pp gap). Catalog captures patterns but misses distributional nuances.
- **Common error tokens**: FP triggers (semicolons 4/4, "willoughby", "marianne"), FN camouflage ("martha" 3/7, "remarked" 2/7, period vocabulary) *almost entirely absent from catalog.*

4.3 Main Insights

1. **Model sophistication beyond vocabulary**: Detectors learn distributional patterns (co-occurrence, positional distributions, register consistency) not captured by human-interpretable AI-isms catalog. 18.3pp accuracy gap between catalog rules and neural model.
2. **FP/FN paradox contradicts intuition**: FPs triggered by formality (punctuation, names) despite *fewer* AI-isms than correct predictions. FNs succeed through minimalism (< 2 AI-isms) + period authenticity. Misclassification driven by register/structure, not vocabulary.
3. **Style mimicry insufficient**: Class 3 reduces AI-isms by 0.76–0.91 but achieves only 0–6.1pp evasion. Detector exploits deeper patterns simple prompting cannot address.
4. **Multiple attribution methods essential**: SHAP highlights contextual/structural markers, Captum identifies character names/formal vocabulary. Complementary perspectives reveal different aspects of model behavior.
5. **Confidence as reliability indicator**: 36.2pp confidence gap between errors (62.4%) and correct predictions (98.6%). Model exhibits epistemic uncertainty, recognizing ambiguous cases.

4.4 Critical Limitations

- **Topic confounding threat**: Detector may exploit 19th-century content rather than authorship signals. Accuracy might drop on modern topics, threatening generalization.
- **Attribution is correlational not causal**: High attribution indicates sensitivity, not necessity. Ablation studies needed but not performed.
- **Model-specific findings**: Results tied to DistilBERT+LoRA and may not generalize across architectures.

5 Task 4: The Turing Test

5.1 Overview

I implemented Genetic Algorithm to evolve AI text until it fools the Try 3 detector (goal: > 90% human probability). I once again used Gemma-3-27B as mutation operator with four strategies (rhythm variation, imperfection injection, casual tone, redundancy addition) and ensemble detector fitness (averaging Try 1, 2, 3 probabilities). After GA success, tested detector on personal writing and discovered fundamental flaw that contradicted Task 2's 99% validation accuracy.

5.2 GA Implementation & Results

Population: 10 diverse initial paragraphs (academic, conversational, technical, persuasive, narrative styles) on modern topic (AI in education).

Selection: Elitism (keep top 3), generate 3 mutations per elite → 9 new candidates per generation.
Mutation strategies:

- **Rhythm**: Vary sentence lengths dramatically
- **Imperfection**: Introduce subtle inconsistencies, archaic words

- **Casual tone**: Add conversational elements, contractions
- **Redundancy**: Natural repetition for emphasis

Results:

- Generation 0: Best = 13.2% human
- Generation 5: Best = 35.9% human
- Generation 10: Best = 78.6% human
- Generation 18: Best = 92.0% human [**GOAL ACHIEVED**]
- Total improvement: +78.8 percentage points

Evolved text characteristics: Archaic vocabulary ("whilst", "endeavour", "hitherto"), formal Victorian prose style, semicolons, subordinate clauses, 19th-century phrasing ("It cannot be denied that...", "One must consider..."). The evolved text reads like Victorian essay prose about 21st-century technology.

5.3 Personal Writing Test

Hypothesis: If detector genuinely learned authorship vs AI generation, my human-written Statement of Purpose should score > 80% human.

Results: **0.9% human (99.1% AI)**

My personal, human-written graduate school SOP flagged with higher AI confidence than most actual AI-generated Class 2 paragraphs in training set.

5.4 What Failed & Why

5.4.1 Failure 1: Manual "Humanization"

Attempt: Added contractions (I've, can't), hedging phrases (perhaps, I think), shortened sentences, casual transitions (Also, Plus).

Result: 0.8% human (−0.1pp, actually *worse*)

Why it failed: Detector doesn't rely on contractions vs formality, sentence length variability, or hedging markers. These are surface features—something deeper at play.

5.4.2 Failure 2: Austen-Style Rewrite

Attempt: Rewrote modern NLP SOP with Victorian vocabulary and 19th-century phrasing ("It is with considerable hope and no small measure of trepidation that I presently submit...").

Result: 1.3% human (+0.4pp, minimal improvement)

Why it failed: Vocabulary alone insufficient—modern NLP topics remain strongly AI-associated despite Victorian style.

5.4.3 Failure 3: Pure 19th-Century Context

Attempt: Completely different SOP as 19th-century governess applying for teaching position, with period-appropriate topics (classical languages, epistolary style, moral instruction).

Result: 8.6% human (+7.7pp, better but still 91.4% AI)

Why it failed: Even authentic Victorian persona on Victorian topics only reached 8.6%. Detector learned era/topic patterns far more deeply than anticipated.

5.5 Systematic Out-of-Vocabulary Testing

Tested **guaranteed human-written** text (pre-LLM era, verified sources) across diverse domains:

Domain	Human Probability
Wikipedia (Bees)	1.9%
Cooking Blog	0.8%
Academic (Climate)	0.9%
Reddit Gaming	8.8%

All modern human writing scored $< 9\%$ **human.**

Control test: Actual Jane Austen (Pride & Prejudice opening): 65.6% human

Pattern: Detector rewards *concentrated* or *prototypical* 19th-century patterns more than actual historical writing with natural variation.

5.6 The Definitive Test: Baby GPT

Discovery: NYT article discussed "Baby GPT"—small AI model trained *exclusively* on Jane Austen novels (no other text).

Hypothesis: If detector learned era/genre rather than authorship, Baby GPT (AI trained on Austen corpus) should score *higher* than modern humans.

Test sample: "You must decide for yourself," said Elizabeth, rather repeatedly; "that is very agreeable displeasure, they will ever be a lively young woman as it will be more disagreeable." (Note grammatical oddities typical of small model trained on limited data—but vocabulary and phrasing authentically Austenian.)

Result: **99.5% human**

Baby GPT scored higher than:

- My actual human writing (0.9%)
- Modern Wikipedia (1.9%)
- Cooking blog (0.8%)
- Academic writing (0.9%)
- Actual Jane Austen text (65.6%)
- GA-evolved text (92.0%)

5.7 Root Cause: Training Data Confounding

The fundamental issue: Task 0 dataset confounded multiple variables:

1. **Historical era** with **authorship**:
    - Human data: 19th-century novels (Austen, Gaskell)
    - AI data: Modern AI models generating 21st-century prose
2. **Literary genre** with **cognitive origin**:
    - Human data: Fiction, narrative, domestic life
    - AI data: Often more expository or neutral when not style-prompted
3. **Vocabulary distribution** with **authorship signal**:
    - Human data: "whilst", "ought", "hitherto", archaic pronouns
    - AI data: "however", "moreover", "can't", modern contractions

What the model actually learned:

$$P(Human \mid Text) \approx P(19th - centuryvocabulary \mid Text)$$

This achieves high validation accuracy *on the training distribution* (all human examples were 19th-century, all AI examples were modern) but does not generalize to modern human writing or AI trained on historical corpora.

5.8 Why Validation Accuracy Was Misleading

Task 2 achieved 99% validation accuracy, suggesting excellent performance. But this measured:

"Can the model distinguish Task 0's Class 1 from Class 2/3?"

Not:

"Can the model distinguish human authorship from AI generation in general?"

The validation set came from the same confounded distribution, so high accuracy did not guarantee true authorship detection.

5.9 How I Understood My Mistake

5.9.1 The GA as Diagnostic Tool

Rather than being a failure, **the GA revealed the detector's true decision boundary**. By optimizing text to maximize human probability, it performed **gradient ascent on the detector's learned features**—exposing that those features were era/genre markers, not authorship signals.

The GA functioned as an adversarial probe:

- It found the shortest path to high fitness
- That path was "add Victorian vocabulary"
- Not "write more authentically human"

5.9.2 Why I Didn't Know Before

1. **Validation metrics looked perfect**: 99% accuracy, high F1 scores, minimal train-test gap. Standard ML evaluation suggested robust learning.
2. **Task 3 interpretability focused on wrong questions**: Analyzed *how* the model distinguished training classes, not *whether* it learned the intended concept. Error analysis examined within-distribution failures, not out-of-distribution generalization.
3. **Implicit assumption of validation set diversity**: Assumed 75-25 split provided sufficient coverage, but both train and validation sampled from same confounded distribution.
4. **Never tested on personal writing**: Most revealing test (absolute ground truth) was never performed until task 4.
5. **Task 0 warning not fully internalized**: Recognized confounding as theoretical risk but didn't test practical consequences until Task 4 adversarial evaluation.

5.10 Comparison with Task 3 Findings

Task 3 interpretation: Class 3 style mimicry achieves only 0–6.1pp evasion (-0.76 to -0.91 AI-isms). Conclusion: Style prompting insufficient, adversarial optimization needed.

Task 4 revised interpretation: GA achieved +78.8pp evasion because it discovered detector's true decision boundary (era/genre) and explicitly optimized for it. Class 3 style prompting only partially effective because:

- Gemma's base training is modern, so "write like Austen" produces modern interpretation of Victorian style
- Prompt added some archaic markers but not enough to fully mimic 19th-century distribution
- GA, through iterative refinement, converged on "pure Victorian prose"

Why Task 3 error analysis missed this: Analyzed semicolons triggering FPs, character names causing confusion, AI-isms catalog having zero overlap with top SHAP tokens—but never tested:

- Out-of-domain modern human text
- Personal human writing
- AI trained on historical corpora

Implicit assumption was validation accuracy (99%) indicated robust learning. Task 4's adversarial approach revealed this assumption was false.

5.11 Error Patterns: Systematic Failure Modes

5.11.1 False Positives: Modern Human → AI

Any text with modern vocabulary, even genuinely human-written, scores < 10% human. Model learned "modern vocabulary = AI" because training AI data was modern.

5.11.2 False Negatives: AI → Human

AI text with 19th-century vocabulary scores > 90% human:

- Baby GPT: 99.5%
- GA-evolved text: 92.0%
- Gemma with Austen-style prompt: ∼85%

Model learned "19th-century vocabulary = human" because training human data was 19th-century.

5.11.3 The Smoking Gun

Baby GPT (99.5%) > Actual Jane Austen (65.6%)

An AI trained on Austen produces *more Austen-like statistical patterns* than Austen herself. Why?

1. Baby GPT trained on multiple Austen novels
2. Learned *average statistical properties* across all novels
3. Generated text is *highly concentrated distillation* of Austen's most frequent patterns
4. Actual Austen includes more variation, rare constructions, idiosyncrasies
5. Detector learned to recognize "prototypical Austen style" (which Baby GPT exemplifies) rather than Austen's full range

Mirrors how GANs can generate "more typical faces" than real faces—they capture the distribution's mode more precisely than any individual sample.

5.12 Key Insights

1. **Adversarial evaluation is necessary, not optional**: Standard validation metrics (accuracy, F1, AUC) can be misleading if validation set shares confounds with training data. Only adversarial testing reveals true model capabilities.
2. **Dataset confounding has catastrophic consequences**: A model cannot learn to distinguish X from Y if all X examples share property A and all Y examples share property B. The model will learn to distinguish A from B instead. In this case: X=Human/Y=AI (intended), A=19th-century/B=Modern (confounded), Model learned A vs B, not X vs Y.
3. **High validation accuracy ≠ robust learning**: 99% accuracy on validation set did not prevent 0.9% human score on my actual writing. Validation set sampled from same distribution as training set; both confounded the same variables.
4. **GAs are powerful diagnostic tools**: Beyond evolving adversarial text, GA revealed detector's decision boundary by finding shortest path to high fitness, exposing that "Victorian vocabulary" was the key feature, not "authentic human writing".
5. **The GA succeeded for the wrong reasons**: Rather than demonstrating detector robustness, it revealed detector limitations. This is a valuable **negative result** that teaches more about ML than a positive result would have.

5.13 How to Fix the Detector (Future Work)

Dataset redesign:

- Diversify human data across eras (19th, 20th, 21st century) and genres (fiction, non-fiction, academic, casual)
- Diversify AI data to match: AI trained on 19th-century corpus (Baby GPT-style), modern corpus (GPT-4, Gemma), varied prompts

- **Critical principle**: For every era/genre in human data, include corresponding AI data. This breaks the confound.

Model architecture:

- Multi-task learning: Primary task (Human vs AI), Auxiliary tasks (Era prediction, Genre prediction) to encourage disentangling factors
- Domain adaptation: Train on multiple domains, use domain-adversarial training for domain-invariant features

Evaluation protocol:

- Out-of-distribution testing: Include unseen eras, genres, authors; personal writing from team members; AI outputs from diverse corpora
- Adversarial robustness testing: Run GA optimization to find adversarial examples; test on Baby GPT-style models; require performance maintenance on adversarial probes

6 Overall Learning Outcomes

1. Dataset construction is iterative hypothesis-driven process with direct downstream implications. Prompt design can dominate model architecture in determining stylistic signals. Dataset confounds persist invisibly until adversarial evaluation. Task 0 confounded era/vocabulary with authorship—undetected through Tasks 1-3 despite 99% validation accuracy.
2. Statistical validation before classifier training essential to identify trivial shortcuts. Effect sizes (Cohen's d) more meaningful than p-values with large samples. Statistical validation within-distribution insufficient. Must test on out-of-distribution data (modern human writing, AI trained on same corpus as human data) to detect confounding.
3. Detection hierarchy reveals signal sources: explicit features (Tier A) brittle but interpretable, semantic content (Tier B) insufficient under topic control, contextual representations (Tier C) capture persistent subtle patterns. Even Tier C's persistent patterns may be confounded. Detector learned era/topic patterns (19th-century vocabulary), not true authorship, despite 99% accuracy and sophisticated contextual modeling.
4. Neural detectors learn beyond human-interpretable markers. Model failure modes (FP formality triggers, FN minimalism evasion) operate on different features than catalog predictions suggest. But "sophisticated learning" can still be spurious correlation. Baby GPT (AI on Austen) scored 99.5% vs actual Austen's 65.6%—detector learned prototypical 19th-century distribution, not cognitive origin.
5. Multiple interpretability methods provide complementary insights. Convergent evidence across SHAP, Captum, and error analysis essential for robust understanding. **Task 4 addition**: Interpretability reveals *how* model distinguishes training classes, not *whether* it learned the intended concept. Task 3 explained decision boundary within confounded distribution but didn't detect the confound itself.
6. Adversarial optimization requires more than avoiding catalog words or style prompting. Systematic approach needed to address distributional patterns and contextual regularities exploited by detectors. But successful adversarial evasion may expose detector flaws rather than demonstrate sophistication. GA revealed detector's true decision boundary (era/vocabulary), diagnosing fundamental limitation.
7. Adversarial evaluation (personal writing, GA optimization, Baby GPT test) is essential to detect confounding that high validation accuracy masks. The most revealing insights came from tests never performed in standard ML pipelines: my own SOP (0.9%), Wikipedia (1.9%), AI trained on Austen (99.5%). Standard metrics (99% accuracy, 36pp confidence gap, zero catalog overlap) suggested robust learning—all misleading.

8. The problem is harder than initially assumed. Modern LLMs can mimic any training distribution (Baby GPT example). Human writing spans enormous stylistic range (Victorian to Reddit). Era/genre confounds are pervasive in available datasets. The field should shift from "detect AI text" to "detect AI text conditioned on era/genre/topic"—acknowledging detection is a distribution-matching problem, not a binary capability gap.