Artificial Text Detection via Magnitude Functions

Selected topics in Data Science, 2025

Hassan Iftikhar Pavel Gurevich

Skoltech

Motivation

Why This Matters LLMs like GPT-4 blur the line between human and AI text. Detection is crucial for trust in academia, media, and online platforms. Existing methods fail as models evolve.

Limitations of Existing Methods

Method	Strengths	Weaknesses
Magnitude Function	- Analyzes geometric shape of word embeddings - Multi-scale detection - Complements other methods	- Computationally heavy - Hard to interpret numerically - Needs good embeddings
Linguistic Feature-Based	- Fast and intuitive - Explains why a text is flagged - Detects simple AI errors	- Fails on sophisticated models - Misses structural cues
Supervised Classification	- High accuracy with good training - Learns new AI behaviors - Detects subtle patterns	- Needs lots of data - May not generalize - Ignores geometry

Our Idea — Magnitude Functions

Use magnitude functions from Topological Data Analysis (TDA).

Text embeddings form a point cloud in high-dimensional space.

Magnitude functions capture the geometry of this cloud at different scales.

Theoretical Background

- Given a metric space A, define:

$$Z_A(a, b) = e^{-d(a,b)}$$

- Magnitude:

$$|A| = \mathbf{1}^T Z_A^{-1} \mathbf{1}$$

- For varying scale t, define:

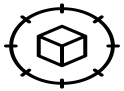
$$\Phi_A(t) = |tA|$$

- This gives a *curve* encoding geometric structure of the text.

Project Pipeline

Data: DAIGT V2 dataset (human + AI texts)

Embeddings: BERT → token vectors



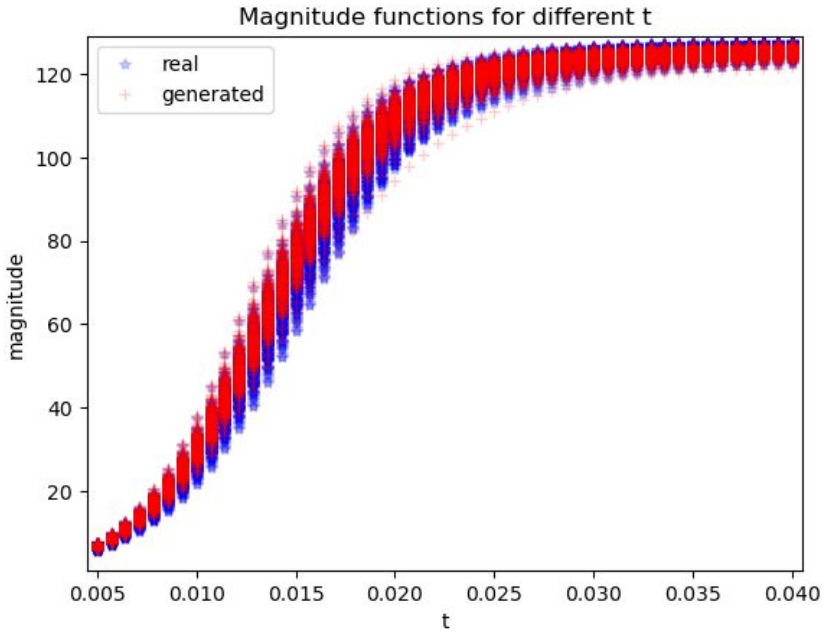
Magnitude Function: For various t values

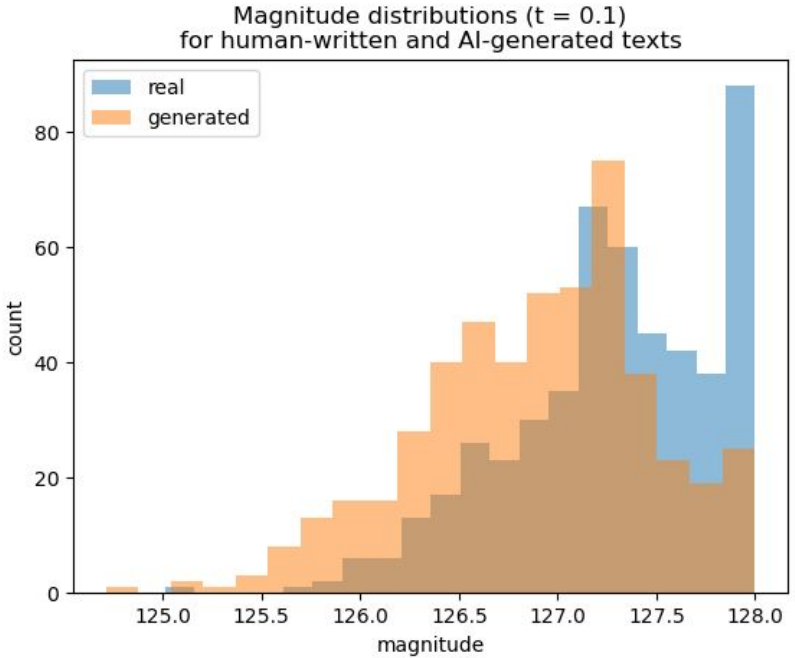
Feature Extraction: Values of $|tA|$

Classification: Logistic Regression (AUC-based evaluation)

Visualizing Magnitude Functions

Observation: Different shapes for real vs. generated texts



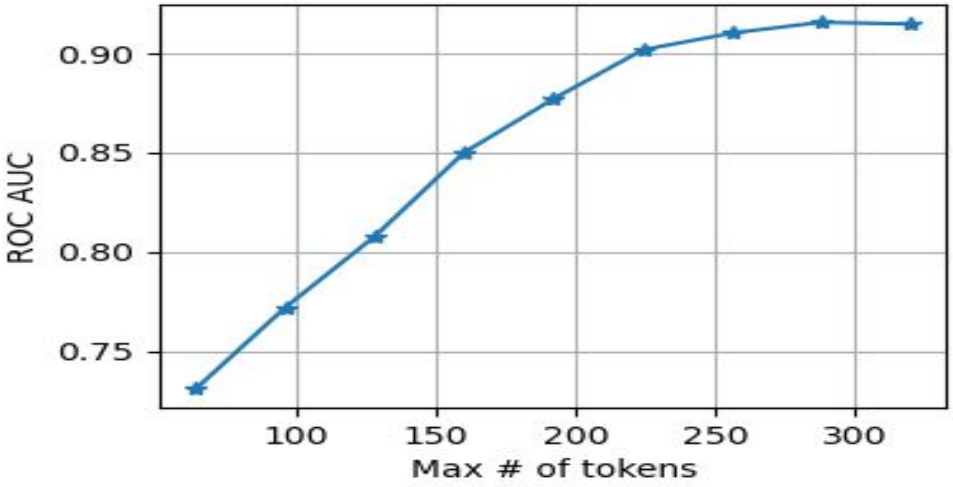


Parameter Optimization

Optimal max tokens: 288

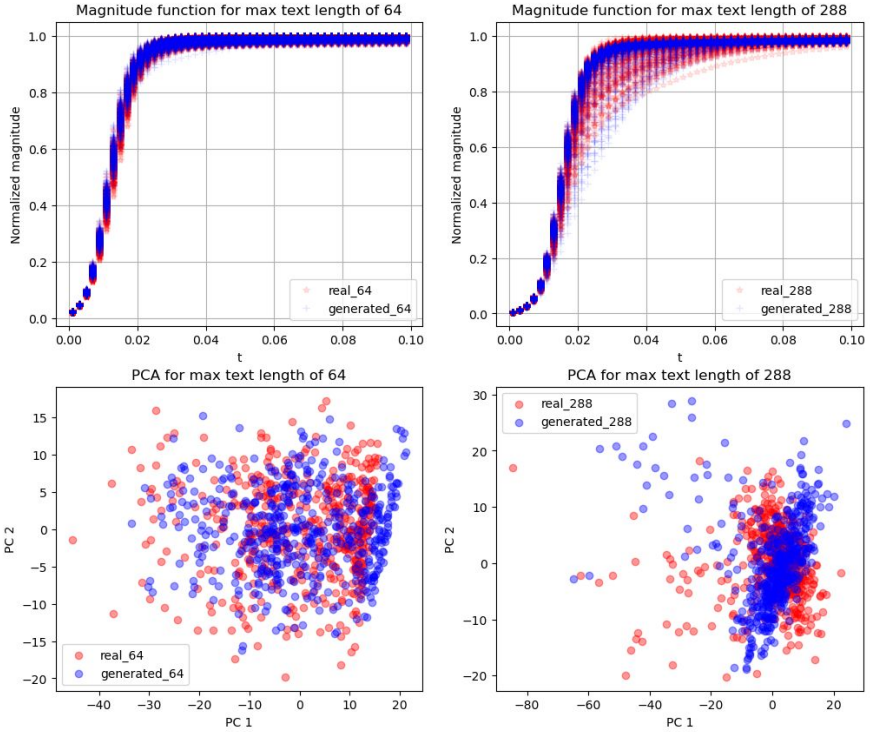
Optimal t range: [1e-8, 0.08]

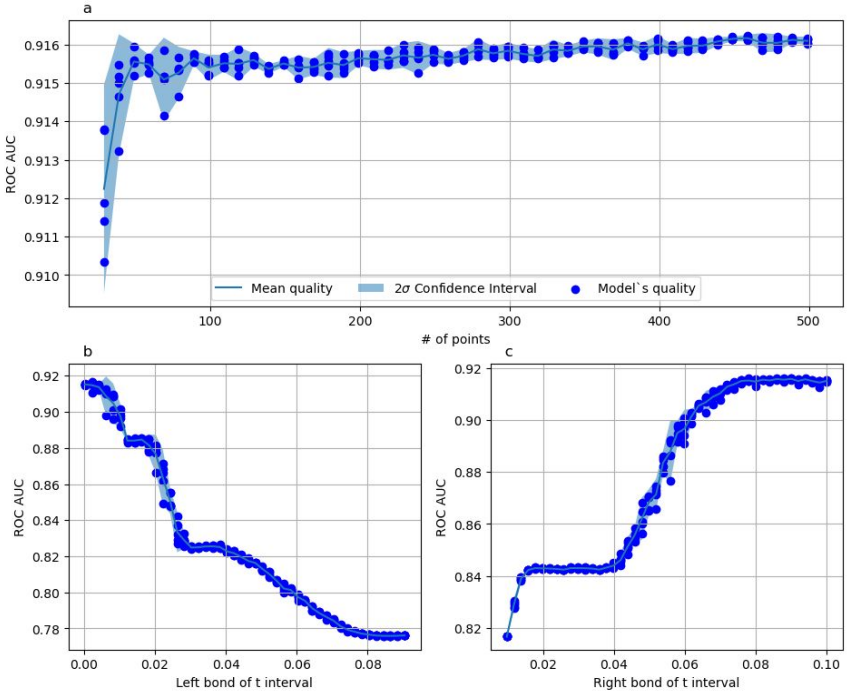
Dense sampling not necessary beyond: ~50 t-points



Some Figures on next slide

Parameter Optimization





Detection Performance

Training data	AUC
Magnitude Only	0.9154
Embedding only	0.9963
Both Combined	**0.9965**

Magnitude adds complementary signal.

Useful as a secondary feature set.

Efficient Computation (CG Solver)

Direct Inversion: $O(n^3)$

Conjugate Gradient: $O(n^2)$, 1.53× faster.

Same approximately accuracy achieved.

Algorithm 1 Conjugate Gradient for Magnitude

Input: Pairwise distances d_{ij},
tolerance ε, max iterations K
Output: Magnitude $|A|$

$Z_{ij} \leftarrow \exp(-d_{ij})$
$b \leftarrow (1, \ldots, 1)^\top$
$w \leftarrow 0, r \leftarrow b, p \leftarrow r, \rho \leftarrow r^\top r, k \leftarrow 0$
repeat
  $q \leftarrow Z \cdot p$
  $\alpha \leftarrow \rho / (p^\top q)$
  $w \leftarrow w + \alpha \cdot p$
  $r \leftarrow r - \alpha \cdot q$
  $\rho_{\text{new}} \leftarrow r^\top r$
  if $\sqrt{\rho_{\text{new}}} \leq \varepsilon \cdot \sqrt{n}$ **then**
    break
  end if
  $\beta \leftarrow \rho_{\text{new}} / \rho$
  $p \leftarrow r + \beta \cdot p$
  $\rho \leftarrow \rho_{\text{new}}$
  $k \leftarrow k + 1$
until $k = K$
return $|A| = \sum_i w_i$

Conclusion

-> Magnitude functions capture topological structure of text embeddings.

-> Boost detection robustness.

-> Efficient CG method enables real-world usage.

Future Directions

-> Nonlinear feature fusion.

-> Magnitude derivatives as features.

-> Robustness to paraphrasing/adversarial attacks.

-> Multi-lingual generalization

-> Benchmarking on the modern LLMs

References

Gehrmann, S., Strobelt, H., Rush, A.
GLTR: Statistical detection and visualization of generated text.
In *Proceedings of the 57th ACL: System Demonstrations*, Florence, 2019.
https://aclanthology.org/P19-3019/

Hu, X., Chen, P.-Y., Ho, T.-Y.
RADAR: Robust AI-text detection via adversarial learning.
In *NeurIPS 2023*.

Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.
Automatic detection of generated text is easiest when humans are fooled.
In *Proceedings of the 58th ACL*, Online, 2020.
https://aclanthology.org/2020.acl-main.164/

Leinster, T., Meckes, M. W.
The magnitude of a metric space: From category theory to geometric measure theory.
In *De Gruyter Open*, 2017.
https://doi.org/10.1515/9783110550832-005

GitHub Repository

🔗 github.com/GurevichPE/Artifical-text-detection-via-magnitude-functions

We'd love your feedback or collaboration!