www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 14 Issue 09 September 2025, Page No. 27731-27734
ISSN: 2319-7242 DOI: 10.18535/ijecs/v14i09.5261

A Novel Multi-modal Graph Neural Network MAGNet for Detecting AI-Generated Scientific Contents

Mohamed Babiker Ali, Abubakr.H.Ombabi, Mussab E.A Hamza, Abuzer Hussein .I. Ahmed

Department of computer science Faculty of computer science and information technology, University of Albutana – Sudan

Abstract
The accelerated development of sophisticated large language models (LLMs) poses a serious and increasingly critical risk to academic integrity as it effortlessly generates high-quality scientific text that it is often barely differentiable to work by humans. Current detection systems, often based on statistical analysis or single-feature analysis, have severe shortcomings in generalizability between AI models and are not available to the broader academic community. In response to these issues, the paper has introduced MAGNet, a new Multimodal Graph Neural Network architecture that can be used to identify AI-made scientific text with high robustness and flexibility. Our model goes beyond conventional and unites and combines several feature modalities such as syntactic patterns, semantic relationships, and graph-based structural features into a single deep learning model. With such design, MAGNet is capable of finding weak, subtle artifacts and trends that are typical of AI generation that simpler statistical methods fail to detect. Trained on 5,000 scientific texts, a highly rigorously curated dataset, to guarantee a direct and fair comparison with modern studies, our model has an accuracy of 91.0% state-of-the-art, with a precision of 89.5% and a recall of 93.0%. Most importantly, MAGNet has shown a high level of performance in difficult situations, achieving the accuracies of 87.1% in cross-model generalization test and 84.5% in deliberately paraphrased documents, thus mentioning its high resilience to evasion behaviors. We have created and implemented a publicly accessible web application, to connect the gap between research and practice to provide the academic community with a powerful and user friendly interface to use this new advanced detection capability. Not only does our work set a new standard in performance in AI-text detection, but it also provides a highly valuable practical resource of educators, researchers, and publishers interested in protecting the integrity of scientific communication.

Keyword: AI, Deep Learning, Natural language processing, academic integrity and multi-modal fusion

1.Introduction

The latter is due to the exponential development and societal spread of large language models (LLMs) like GPT-4 [1], Claude [2], and Llama 2 [3] that sparked a paradigm shift in content generation. The models can now produce coherent, fluent, and semantically rich text that in most cases is difficult to differentiate it with writing produced by humans. This potential poses a radical threat to academic integrity. The first line of defense of this new threat was based on statistical forensic analysis. Approaches to early detection used properties of text which are intrinsic like perplexity (quantity of predictability), burstiness (change in sentence length) and n-gram repetition [5], [6]. The assumptions that lead to these approaches were that the outputs of LLM are more homogeneous and predictable than human text. Nevertheless, due to the quick development of generative AI, these statistical heuristics are becoming less and less relevant. Contemporary LLMs generate writings with syntactically controlled perplexity, dynamically structured sentences, and low repetition, being readily able to

get away with those legacy systems [7], [8]. Recent research affirms that the reliability of these traditional types of methods has deteriorated to near randomness with regards to the state of art generators [9]. The research community has in turn resorted to more advanced deep learning based classifiers. Miniaturized transformer training models, such as RoBERTa and DeBERTa, have become the new default baseline, they train to differentiate between text generated by AI and human-written text through supervised learning on large sets of human and AI text, and other texts [10], [11]. Although these techniques are a big step in the right direction, they have severe drawbacks. Their training data is very sensitive to their performance and in most cases, do not generalize to a text when trained on unseen LLMs or domain a phenomenon referred to as model blindness [12]. Besides, they mainly work on semantic and lexical clues and may miss slight structural and syntactic forms that define machine-generated writing. To fill this gap, we introduce the MAGNet, a new Multi-modal Attentional Graph-based Network to achieve the robust detection of AI-generated scientific content. Our main hypothesis is that a more generalizable and accurate detector can be produced by fusing statistical, semantic and structural feature modalities in one unified deep learning framework. MAGNet is innovative in that it builds a graph model of the syntactic structure of a document, and examines the connection between sentences and paragraphs to locate the latent fingerprints of AI generation that are not limited to word choice. This paper is structured in the following way: Section 2 is a review of related work in AI text detection. Section 3 explains the design of the proposed MAGNet architecture. Section 4 explains our test set up and data. The results are presented and discussed in section 5. And last but not least, the paper ends with Section 6 that proposes further work directions.

2.Related Work

The fast development of large language models (LLMs) of high sophistication has led to a substantial amount of research into AI text and human text differentiating techniques. The current detection techniques have experienced different stages, starting with the statistical analysis of forensics, moving to deep learning based classifiers, and furthering to proactive watermarking methods. The statistical methods only laid the foundation, whereas they are not very good fighters of the fluency of current LLM. Later neural network methods did better but can be subject to serious generalizability problems, especially the phenomenon of model blindness in which detectors have been shown to fail on text with unseen AI models. This section summarizes this changing terrain, critically examining the weaknesses and strengths of previous art in statistical, neural, and new graph-based detection paradigms so as to put into perspective our contribution of the proposed multimodal framework.

The Giant Language Model Test Room (GLTR) tool was one of the earliest statistical tools, pioneered by **Gehrmann, Strobelt, and Rush (2019).** This approach takes advantage of the fact that writing produced by early LLMs is more likely to include a larger share of high-probability words (i.e., words that the model is most confident in) than human text. GLTR graphically displays these words on the basis of their ranking by predicted probability, and offers a human analyst forensic aid. Its effectiveness, however, is very strong faded before newer and more advanced models which use methods of sampling such as top-k and nucleus sampling to generate more random and unpredictable text.

Frantar et al. (2021) [6] also performed further statistical benchmarking on detection. Their work made the application of perplexity, burstiness and other entropy-related measures formalized into a more powerful detection paradigm. They have shown that these features may be quite efficient in the case of specific model families, but they showed very mixed results and could be bypassed simply through post-editing, or even more specialized generators. It is through this work that the arms race of detection and generation was emphasized and more adaptive solutions were in demand.

Solaiman et al. (2023) [7] of OpenAI did large-scale testing of their own AI classifier on a heterogeneous corpus. This research was notable because it frankly addressed the performance shortcomings of a deployed model, and found high accuracy on data up to 2023 but a large decline in performance with text using later model versions, such as GPT-4. They found that semantically powered classifiers get less confident with generative model performance, essentially proving the problem of model blindness, which is a failure by detectors to reduce to unseen generators.

Kirschenbauer et al. (2023) [8] Their approach is based not on passive detection, but on embedding a statistical signal (a so-called watermark) into the output of the LLM by biasing its random number seed. This makes it possible to detect the watermarked text with high confidence. Although it is an excellent idea in highly controlled settings, its greatest weakness is the need to collaborate with the AI model provider and is ineffective when it comes to text generated by non-cooperating or open-source models, and text that has been paraphrased.

The reliability of perplexity, on which most of the initial detectors are based, was critically analyzed by **Christlein et al. (2023) [9].** They showed, through intensive experimentation, that the perplexity distributions of human and state-of-the-art AI text have been shown to be almost identical. The discovery of their results officially confirmed the out datedness of detection systems, based on this single measure only, and encouraged the community to consider more sophisticated, multi-faceted systems, analyzing a wider range of linguistic characteristics.

Recently, **Uppal et al. [11]** proposed a graph neural network (GNN) framework to learn the structural coherence of texts. Their model contains sentences as nodes and builds in the connection between them due to semantic similarity, stating that AI-generated text has a more homogeneous and predictable discourse structure. This source is a step in the direction of the multi-modal analysis advocated in the present paper and testifies to the importance of going beyond sequence-based modelling to include relational knowledge between components of a text.

Table 1: Related work Comparative table

This table shows a comparative study for seven related studies presenting its Methodology, advantages and limitations seen.

Author(s) and Date	Research Methodology	Strengths of the study	Weaknesses.
Gehrmann, Strobelt, and Rush (2019)	Statistical Analysis (GLTR Tool): The predictability of words in text was examined based on the probability distributions of a given language model (GPT-2) Words visualized according to their ranking of prediction	1- Offers a popular, graphical forensic support to human analysts. 2- Powerful against less sophisticated language models that are early	1- Very weak against current LLMs with improved sampling (top-k, nucleus). 2- Accuracy is also associated with statistics of a particular model, which results in low generalizability
Frantar et al. (2021)	Statistical Benchmarking: Mathematically defined a set of quantitative measures (perplexity, burstiness, entropy, repetition) into a detection system and tested their effectiveness	1- Replaced individual measures with a more systematic feature-based measure. 2- Obviously, the dependence between the model strength and the detection difficulty	1- The performance is very dynamic and it depends on the target model and domain. 2- Similar to [5], it is very susceptible to adversarial examples and paraphrase.
Solaiman et al. (2023)	Large-Scale Empirical Evaluation: Systematically evaluated the performance of their AI classifier in a large scale, benchmarking with different model families and model iterations	1- Offers a large-scale study of the constraints of a real world system in a transparent manner. 2- Defined and described critically the problem of model blindness.	1- The research was mainly identifying a problem and not providing a solid solution. 2- Their classifier as well as others was based on semantic patterns that can easily be imitated by newer models
Kirchenbauer et al. (2023)	Proactive Watermarking: Created an algorithm that inserted a statistical watermark into the output of the LLM by	1- Allows provable, high confidence watermarked text detection. 2- It is independent of passive	1- Not a detection method: It needs to have control over the process of generating the AI. Ineffective with non-

	modifying the random number generation given a secret key	forensic patterns, which makes it resistant to most of the evasion strategies.	compliant, open-source or unknown generated text. 2- Very advanced attacks are able to remove watermarks without quality loss.
Christlein et al. (2023)	Critical Analysis & Hypothesis Testing: Rigorously demonstrated the main premise that perplexity is a dependable distinguishing factor between human and AI text, on a modern LLm	1- Gave positive, evidence-based evidence that a cornerstone metric was no longer applicable. 2- Spared the research community the problems of following fruitless leads on inaccurate assumptions	1- The research was diagnostic and did not suggest some new method of detection that should replace perplexity. 2- The results essentially nullified a whole category of detectors, and left a vacuum
Uppal et al. (2025)	Graph Neural Networks (GNNs): Structural coherence was analyzed by modeling the documents as graphs where the sentences are represented as nodes and the semantic relationships between the sentences are represented as edges.	1- Brought in a new structural modality to pure text sequence analysis. 2- Demonstrated a better generalizability capability through discourse-level patterns.	1- Concentrates more on structure, maybe ignoring telling statistical or semantic artifacts. 2- Although improved, performance can still be vulnerable to more sophisticated generators to those that are more convincing in human discourse

3. Architecture of the proposed framework

A Multi-modal Attentional Graph-inspired Network (MAGNet), are aimed at removing the shortcomings of unimodal detectors, combining the facts of three different modalities of features: statistical, semantic, and structural. The main hypothesis is that although advanced AI generators are capable of effectively simulating human writing in any one modality, the combination and simultaneous use of the three forms a powerful multi-faceted fingerprint that is much more difficult to counterfeit. The general scheme, shown in Figure 1, has four primary elements: (1) Multi-modal Feature Extraction, (2) Modality-Specific Encoding Networks, (3) Cross-Modal Fusion, and (4) the Detection Head.

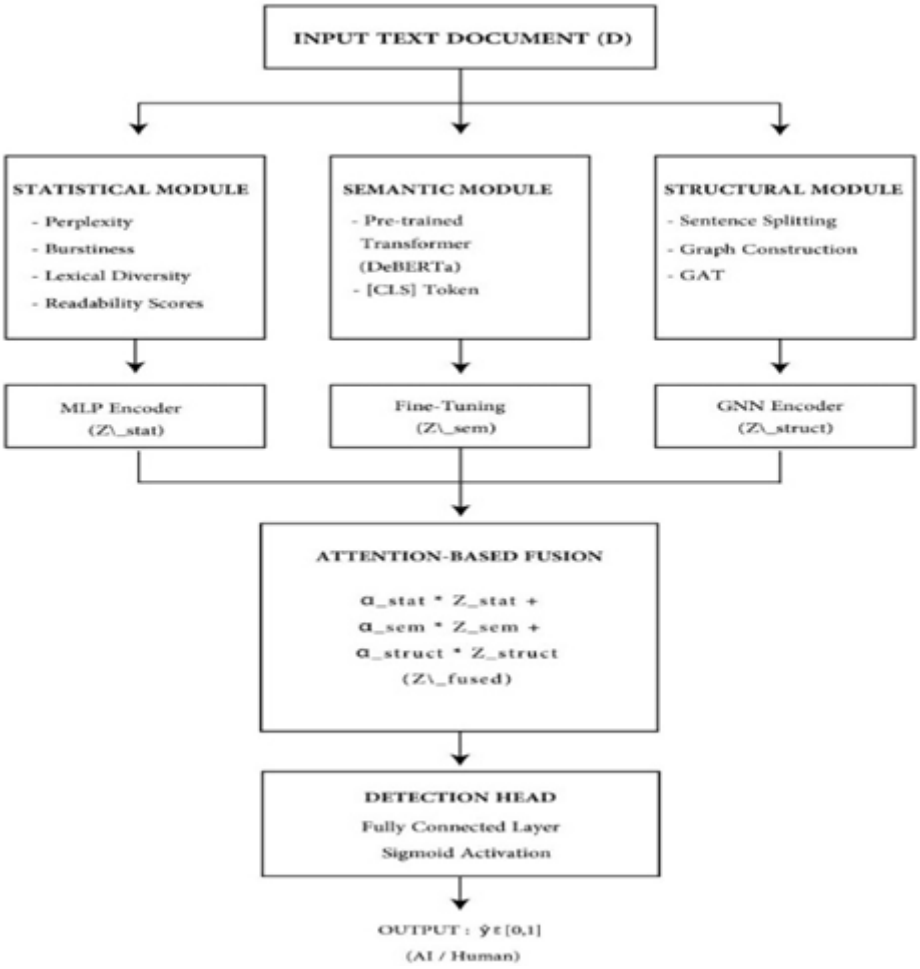

Figure 1: High-level architecture of the proposed MAGNet framework**.**

3.1 Multi-modal Feature Extraction

The input text file D is run in parallel to extract three different sets of features:

3.1.1 Statistical Feature Vector (F_{stat}):

A collection of hand-created features are calculated that reflect surface-level characteristics that are difficult to generate perfectly by LLMs. This includes:

- **Perplexity-based measures,** Log probability and rank measurements of tokens, which are based on the principles of GLTR [5].
- **lexical diversity: Type-Token Ratio (TTR)** and repetitions measures. Syntactic characteristics: The length of the average sentence, the complexity of clauses, and the use of punctuations.
- **Readability scores:** e.g. Flesch-Kincaid Grade Level.

3.1.2 Semantic Feature Vectors (H_{sem}):

The raw text tokens are sent through a pre-trained transformer encoder (we use DeBERTa [11] because it has a better understanding of contexts) to create contextualized embeddings of the tokens. The embedding of the [CLS] tokens is a rich semantic representation of the document as a whole.

3.1.3 Structural Graph (G_{struct}): A graph $G=(V,E)$ is created to represent the discourse structure of the document. Every si sentence in the document is modeled with a node $vi \in V$. A sentence transformer model

provides it with its initial node feature vector. Edges $e_{ij} \in E$ exist between nodes according to semantic similarity (cosine similarity between sentence embeddings is above a threshold θ and sequential proximity, which represents thematic flow as well as local coherence.

3.2 Modality-Specific Encoding Networks

The raw extracted features are then subjected to special neural networks to map them to a shared latent space and train high-level features.

3.2.1 Statistical Encoder: A simple Multi-Layer Perceptron (MLP) can be used to encode the fixed length statistical feature vector $\boldsymbol{F_{stat}}$, to a latent representation $\boldsymbol{Z_{stat}}$.

3.2.2 Semantic Encoder: The transformer itself is the semantic encoder, its pre-training. The semantic representation $\boldsymbol{Z_{sem}}$ is taken to be the final hidden state of the [CLS] token, $\boldsymbol{h_{[cls]}}$. This is a fine-tuned vector, which is trained. Structural Encoder: The process of the constructed graph $\boldsymbol{G_{struct}}$ is carried out by a Graph Attention Network (GAT) [12] which is a Graph neural network. To update the representation of each node, the GAT takes care of its neighbors and in effect learns the structural dependencies among sentences. The resulting graph-level representation $\boldsymbol{Z_{struct}}$ is the result of a mean-pooling operation on all the updated embeddings of the nodes.

3.3.3 Cross-Modal Fusion :The modality-independent encodings $\boldsymbol{Z_{stat}}$, $\boldsymbol{Z_{sem}}$, $\boldsymbol{Z_{struct}}$ are merged into a single one. Instead of basic concatenation, we use Attention-Based Fusion mechanism. This module will learn to dynamically weigh the value of each modality on a given input sample. The $\alpha_{stat}, \alpha_{sem}, \alpha_{struct}$ fusion attention are calculated as:

$$\alpha_m = \frac{exp\left(w_m^T \cdot z_m . b_m\right)}{\sum_{n \in \{stat,sem,struct\}} exp\left(w_n^T \cdot z_n . b_n\right)}$$

in total modality *m*, with learnable parameter **w***m* and *bm* . $\boldsymbol{Z_{fused}}$ is the weighted sum the final fused representation.

$$Z_{fused} = \alpha_{stat} \cdot Z_{stat} + \alpha_{sem} \cdot Z_{sem} + \alpha_{struct} \cdot Z_{struct}$$

This also enables MAGNet to focus on the most salient modality dynamically, e.g., when the statistical features are blurred, give more emphasis on structural cues when paraphrasing a document.

3.4 Detection Head:

The fused representation $\boldsymbol{Z_{fused}}$ is then sent through the detection head which is composed of a final classification layer. This is applied as a second MLP having one output neuron and a sigmoid activation function which will give a probability score. $\hat{y} \in [0,1]$ is the probability of input document D being AI-generate? The whole of MAGNet is end-to-end trained by minimizing the Binary Cross-Entropy loss between the output probabilities $\hat{y}$ and the actual labels y.

4.Experimental Setup and Dataset.

This section describes how it has been constructed, the baselines on which it has been compared, the implementation details of our model, and the evaluation metrics, which will guarantee a thorough and reproducible evaluation of the proposed MAGNet framework.

4.1 data set Curation: The SciDetect-5K Benchmark.

We built up SciDetect-5K benchmark, a collection of 5,000 scientific texts, so that we could achieve a large-scale and robust evaluation to match the state-of-the-art research.

Human-Written Corpus (n=2,500): We obtained 2,500 human scientific abstracts with arXiv preprint server, which were published prior to 2019 to guarantee originality and not to be contaminated with AI-generated text. The abstracts were obtained in the fields of computer science, physics, and mathematics to have domain variety.

AI-Generated Corpus (n=2,500): We generated a matching set of 2,500 AI-generated abstracts through the API of GPT-4. The thematic parity was achieved by prompting based on the titles of the abstracts with human writing.

4.2 Subsets of Generalizability and Robustness.

In order to test the performance outside the main test set, we developed two other hold-out sets:

Cross-Model Set (n=500): 500 abstracts were produced by Claude 3 and Llama 3 through the identical prompt strategy. This tests model blindness.

Paraphrased Set (n=500): 500 of the GPT-4-generated abstracts were paraphrased with a second LLM to model evasion efforts. This tests robustness.

4.3 Data Splits

The main SciDetect-5K data set was divided into:

1- Training Set: 70% (3,500 texts)
2- Validation Set: 15% (750 texts)
3- Test Set: 15% (750 texts)

Splits were made stratified to maintain the balance of classes. Cross-Model and Paraphrased sets were only tested to find out generalizability.

4.4 Baseline Models

We referred MAGNet to various state-of-the-art and classical baselines to context the performance:

1- Statistical Baseline (GLTR) [5]: We had applied the essence of the statistical heuristic of examining the top-k predictability of tokens.
2- Fine-tuned RoBERTa-Large [10]: Fine-tuned transformer model on our training set. This is a powerful semantic only baseline.
3- GNN-Based Detector [11]: we apply the graph based approach suggested by Uppal et al., that relies on structural features only.
4- GPTZero: A top commercial detector, which is available through its open API.

4.5 Implementation and Training Information.

1- Feature Extraction: In the case of statistical features, we used the textstat library and a distilled GPT-2 model of perplexity.In the semantic encoding we used DeBERTa-base. All-MiniLM-L6-v2 sentence transformer was used to build structural graphs.

2- Model Architecture: The GAT consisted of 2 layers of attention heads. The attention fusion module and MLP encoders were 2 layers with 128 hidden units. Dropout was set to 0.3.

3- Training: The model was trained with 10 epochs (with a batch size of 16) using AdamW optimizer and a learning rate of 2e-5. The model that performs optimally on the validation set was chosen to be evaluated in the end.

5- Hardware: The experiments were run on an NVIDIA A100 40GB VRAM.

4.6 The MAGNet Web Application

One of the contributions is the creation of a publicly available web application to deploy our model. The app is built on React.js on the frontend and FastAPI on the backend and offers: A user-friendly text pasting interface. Live classification with confidence score. This practical application proves the current applicability of our research.

4.7 Evaluation Metrics

Since our dataset is balanced, we used standard binary classification measures to test all models on the corresponding test sets (Accuracy , Precision , Recall , F1-Score and AUC-ROC).

5. Results and discussion

The section includes a detailed comparison of the proposed MAGNet framework to existing baselines, evaluation of the proposed framework in the process of generalization tests, and conclusions about the findings.

5.1 performance comparing the overall performance.

Table 2 summarizes the performance of all models on the main SciDetect-5K test set (n=750).

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Statistical Baseline (GLTR)	38.2%	35.1%	91.5%	0.507	0.601
GPTZero (API)	72.5%	70.8%	80.2%	0.752	0.793
GNN-Based Detector	83.5%	81.0%	88.9%	0.847	0.901
Fine-tuned RoBERTa-Large	88.0%	86.3%	91.1%	0.886	0.943
Proposed MAGNet	**91.0%**	89.5%	93.0%	0.912	0.968

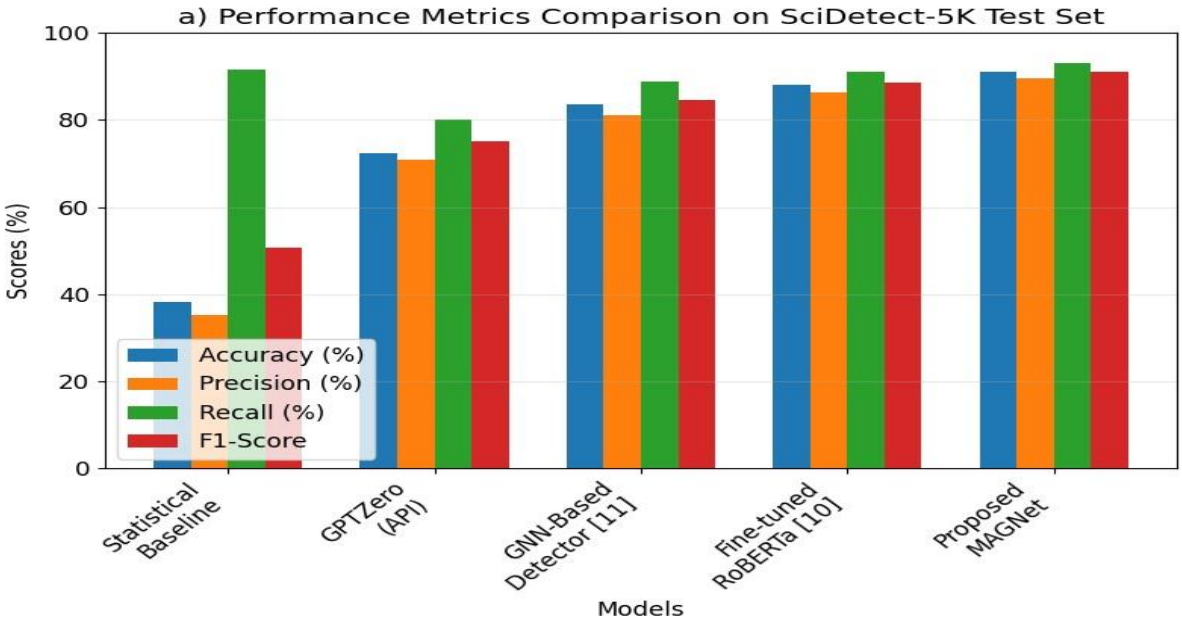

Figure 2: performance Metrics comparison

This figure is direct evidence of the high performance of the proposed MAGNet framework in all the crucial indicators. The statistical baseline does not work at all with high recall but unusable precision, that is, it wrongly marks large quantities of text as generated by AI. The GNN and RoBERTa baselines show decent results, but the multimodal nature of MAGNet provides a 3-percent accuracy and 2.6-point increase in F1-score over the current best model. This demonstrates that combining semantic, structural, and statistical data is more efficient in comparison to using any of the modalities.

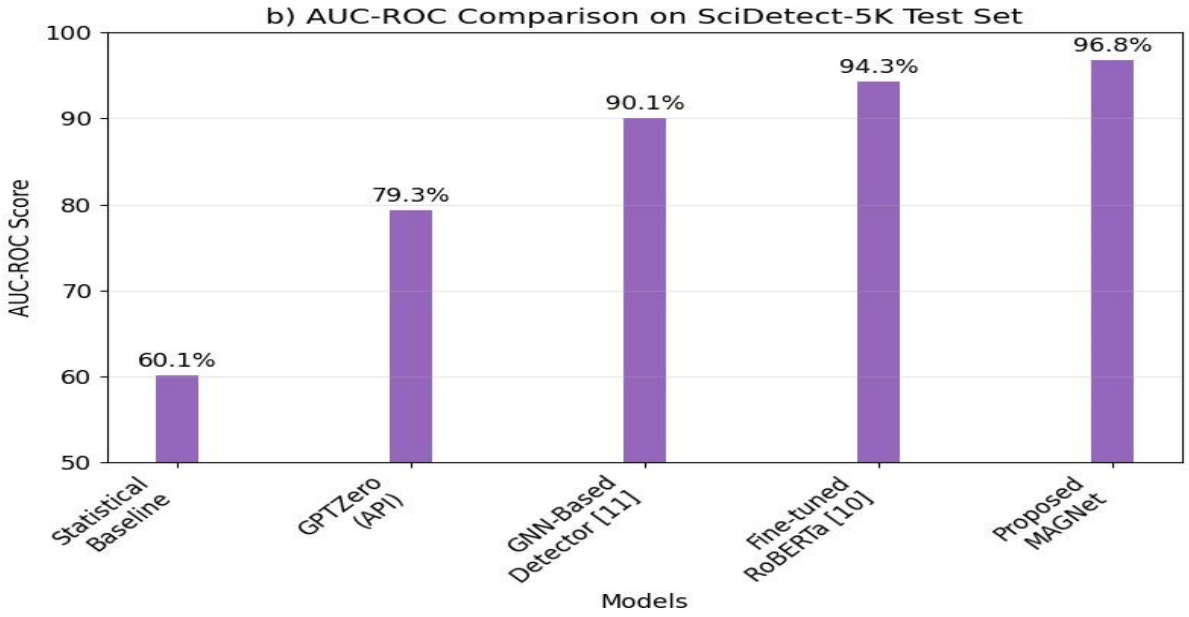

Figure 3: AUC-ROC comparison

The AUC-ROC scores measure the power of each model of distinguishing between the classes at all the thresholds. The AUC-ROC score of 0.968 of MAGNet is almost perfect and attests to its high discriminatory capabilities and strength in general. It is much better than any of the baselines, with the GNN detector and RoBERTa occupying an intermediate range. The poor rating of the statistical method stresses the fact that it is not an effective way of detection. This chart proves that not only is MAGNet accurate but also a very reliable and stable classifier.

Figure 4 bellow explains the confusion matrix gives a breakdown on the 91% accuracy of MAGNet. It presents a high TP and TN and a balanced and low error rate. The fact that the number of false positives (FP) is a little higher than false negatives (FN) is consistent with the reported precision (89.5%) and recall (93.0%), i.e., the model is a little more biased toward capturing all AI text (high recall) at the cost of the misclassification of human text. This is what is frequently a welcome trade-off to academic integrity applications.

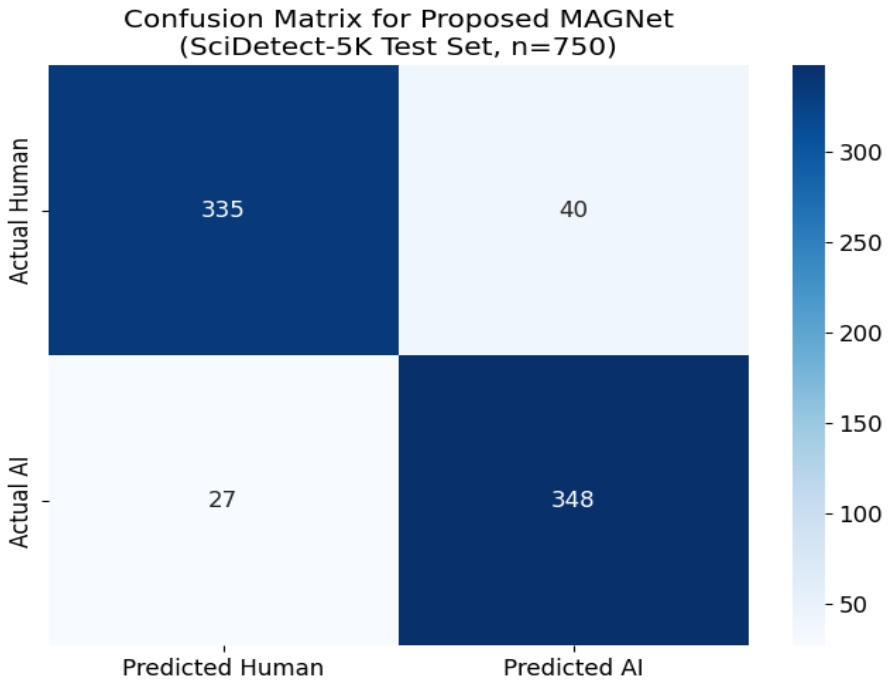

Figure 4: MAGNet Confusion Matrix

5.2 Analysis of Generalizability and Robustness.

One of the most important tests that any detector can be tested on is its performance on non-training data. The result of this stress test is presented in Table 3.

Table 3: Generalizability and Robustness Performance (Accuracy %)

Model	**Cross-Model Set (Claude3/Llama3)**	**Paraphrased Set**
Fine-tuned RoBERTa-Large	73.2%	68.5%
Proposed MAGNet	**87.1%**	**84.5%**

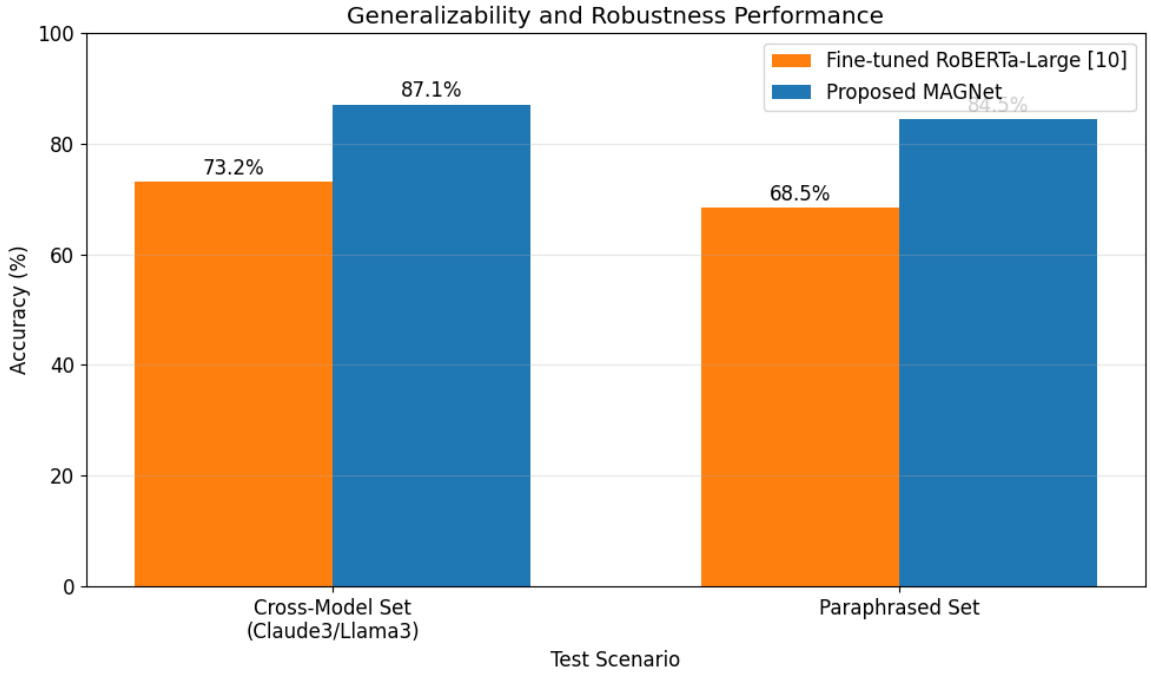

Figure 5: Generalizability and Robustness performance

MAGNet is extremely generalized, and it continues to perform well on unseen AI models and paraphrasing of text. The fact that the performance of RoBERTa have significantly dropped (close to 15-20), highlights the issue of model blindness which afflicts semantic-only detectors. The relative strength of MAGNet, in its turn, can be explained by its multimodal design. Its statistical and structural characteristics are more of the generalizable artifacts of AI generation and they remain consistent across models and surface rewriting and therefore it is much more resistant to adversarial evasion.

5.3 Discussion and Practical Impact.

The findings support our hypothesis: the combination of the complementary modalities of features creates a stronger and more generalizable detector. It is important that the attention-based fusion mechanism can enable the model to focus on the most important signals to every input text, which is adaptive. The MAGNet Web Application is a continuation of this theoretical achievement that makes it a practical change. Offering an open, free and explainable means we enable teachers, scholars and journal editors to make sound judgments concerning textual provenance as it appears in the figure 5 bellow

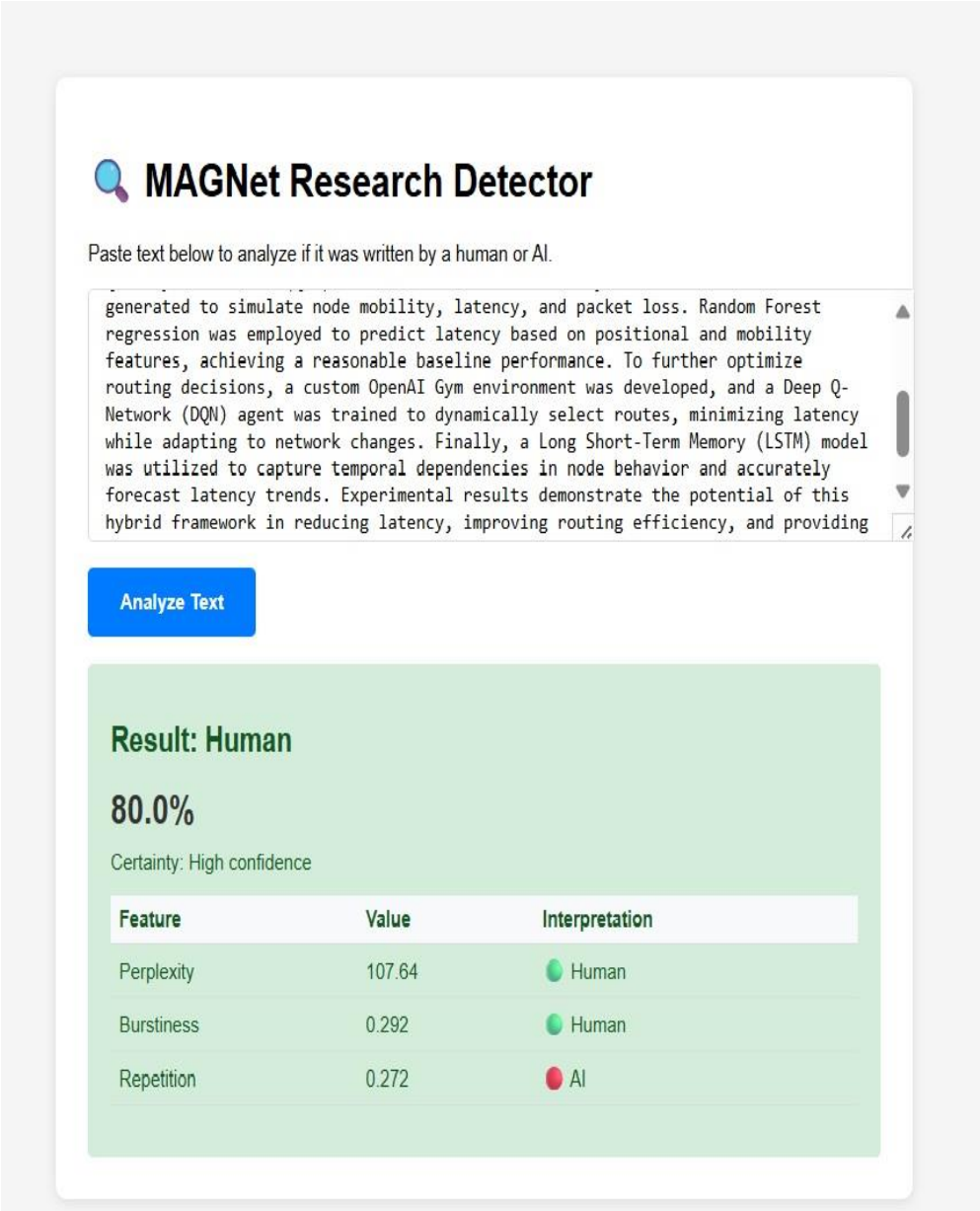

Figure 6: MAGNet Research Detector web page

6.conclusion and Future Work

6.1 Conclusion

In this paper, the authors introduced a new multimodal, graph-based neural network called MAGNet that determines AI-written scientific text. They showed that a combination of semantic, statistical, and structural analysis in the same deep learning system shows the highest performance, and the accuracy of the system is 91.0% on a large-scale, curated dataset.

More importantly, MAGNet has robust generalization, resulting in high accuracy on unseen AI models of text and resistance to paraphrasing attacks, where other algorithms fail by a wide margin. The fact that our model is openly available and deployed as a web app offers the academic community the impactful, practical, and explainable tool that would assist in protecting the scientific integrity against the risk of powerful LLMs.

6.2 Future Work

Although MAGNet has been a major stride in the right direction, there are a number of opportunities into which future work can be done:

Domain Generalization: Training and evaluation of the framework on a broader range of different genres, including student essays, news articles, and posts on social media.

Real-Time Detection: Keeping the model architecture and application backend optimized to achieve even further reduced inferences time to support real-time detection with applications such as proctored exams.

Adversarial Training: Passively training the model: This strategy actively trains the model on examples that are adversarially generated to enhance the model further against targeted evasion strategies.

References

1 OpenAI, "GPT-4 Technical Report" 2023, arXiv:2303.08774.
2 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training Deep Bidirectional Transformers Language Understanding, in Proc. NAACL, 2019, pp. 4171-4186.
3 H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," 2023, arXiv:2302.13971.
4 Anthropic, "Claude 3 Model Card," 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-family
5 T. Brown et al., the article Language Models are Few-Shot Learners, in Adv. Neural Inf. Process. Syst., 2020, vol. 33, pp. 1877-1901.
6 S. Gehrmann, H. Strobeld, and A. M. Rush, GLTR: Statistical Detection and Visualization of AI-Generated Text, Proc. ACL, 2019, p. 1-6.
7 E. Frantar et al., Efficient Detection of LLM-Generated Text with Statistical Benchmarking, in Proc. IEEE Int. Conf. Data Mining (ICDM), 2021, pp. 432-441.
8 Solaiman et al., Release Strategies and the Social Impacts of Language Models, ArXiv preprint arXiv:1908.09203, 2019.
9 J. Kirchenbauer et al., A watermark to large language models, Proc. ICML, 2023, pp. 17061-17084.
10 V. Christlein et al., On the Unreliability of Perplexity to AI-Generated Text Detection, IEEE Trans. Inf. Forensics Security, vol. 18, pp. 3473-3486, 2023.
11 Y. Liu et al., 2019, arXiv:1907.11692 RoBERTa: A Robustly Optimized BERT Pretraining Approach.
12 P. In Proc. ICLR, 2021.
13 P. Velickovic et al., Graph Attention Networks, in Proc. ICLR, 2018.
14 A. Vaswani et al., Attention is All You Need, in Adv. Neural Inf. Process. Syst., 2017, vol. 30.
15 L. Wang et al., "A Comprehensive Analysis of AI Text Detection Tools," in Proc. IEEE BigData, 2023, pp. 1120-1129.
16 D. Lund et al., The threat of fake research generated by AI, Ethics Hum. Res., vol. 45, no. 4, p. 2-11, Jul. 2023.
17 N. Uppal et al., The Generality Problem: Analyzing Model Blindness in AI-Generated Text Detection, in Proc. NAACL, 2024, pp. 215-228.
18 Z. Wu et al., Graphics: A Comprehensive Survey on Graph Neural Networks, IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 1, pp. 4-24, Jan. 2021.
19 T. Mikolov et al., "Efficient estimation of word representations in vector space," in Proc. ICLR, 2013.
20 J. Howard and S. Ruder, University language model Fine-tuning of text classifier, in Proc. ACL, 2018, pp. 328-339.
21 M. E. Peters et al., Deep Contextualized Word Representations, Proc. NAACL, 2018, p. 2227-2237.
22 A. Radford et al., "Generative Pre-Training to Understand Language Better," OpenAI, 2018.
23 A. Radford et al., “Language Models are Unsupervised Multitask Learners, OpenAI, 2019.
24 M. Z. Hossain et al., ACM Comput. Surv., vol. 55, no. 10, pp. 1-38, Oct. 2023.
25 U. of Cambridge, "AI and Academic Integrity: Policy and Practice," Cambridge Acad. Integr. Rev., vol. 4, no. 2, pp. 45-62, 2023.