UNCERTAINTY AWARE PARAPHRASE PLAGIARISM DETECTION

Taejin Park, Gayatri Susarla
Department of Applied Mathematics & Statistics, Data Science
Stony Brook University
taejin.park.2@stonybrook.edu,gayatri.susarla@stonybrook.edu

ABSTRACT

Large language models (LLMs) can rapidly re-paraphrase source text in ways that evade traditional string-matching plagiarism checkers. This project investigates the dual problem of (i) generating machine paraphrases with a lightweight LLM and (ii) detecting such paraphrases using both conventional services and fine-tuned Transformer classifiers. Building on recent research on machine-paraphrased plagiarism, watermarking for LLMs, and paraphrase detection, we fine-tune T5-small to create paraphrases over the 262k-example Autoregressive Paraphrase Dataset, and train a RoBERTa-base model to differentiate original from machine-generated text. Our experiments show that the Transformer detector reaches an F-1 score of 0.989, outperforming commercial plagiarism checkers by an order of magnitude, while preserving human-perceived text quality. We did an analysis on the model with and without a watermarked sample and evaluated some parameter metrics. Our code is available here https://github.com/GAYATRI-SIVANI-SUSARLA/LLM_PROJECT

1 INTRODUCTION

The rapid growth of generative LLMs such as GPT-3, GPT-4, and open-source models now enables students, researchers, and malicious actors to automatically rewrite arbitrary text at near-human quality. These machine-paraphrased passages often escape detection by surface-level similarity metrics used in student writing analytics and academic integrity software. This project addresses the following questions:

1. **Generation**: How well can a compact open-source model (T5-small) generate high-quality paraphrases on academic prose?
2. **Detection**: Can a fine-tuned Transformer outperform commercial plagiarism detectors on the same inputs?
3. **Forensics**: How do current LLM watermarking techniques integrate with a detection workflow?

2 RELATED WORK

Large-language-model paraphrasing has been shown to slash human and algorithmic plagiarism-detection accuracy - GPT-3 / T5 paraphrases cut it to roughly 53%, and the best traditional detector stalls at F-1 score of 0.66 [1]. To curb such misuse, Kirchenbauer et al. proposed a lightweight statistical watermark that can be spotted from plain text with minimal impact on fluency [2]; however, it offers no protection when un-watermarked models or post-edited paraphrases are used. Baseline detectors that couple word-embeddings with shallow classifiers reach only F-1 score of 0.72 on academic corpora, while commercial string-matching services fare even worse [3]. Progress has accelerated with the release of the 262k - example Autoregressive Paraphrase Dataset (APD) [4], which supports training at scale, and with Transformer-based detectors that markedly outperform legacy tools, as shown in recent coursework evaluations [5]. Building on this line of work, our study leverages APD to fine-tune a lightweight T5-small paraphraser and trains a RoBERTa-base classifier over commercial checkers—and we quantify how its judgments shift in the presence or absence of the watermark, highlighting the complementarity of watermarking and neural detection.

Topic	**Key Findings**	**Source**
Machine-paraphrased plagiarism	GPT-3 and T5 paraphrases reduce human detection accuracy to 53 %; best machine detector F1 ≈ 0.66	Wahle *et al.*, EMNLP 2022
Watermarking LLM outputs	Statistical "green-list/red-list" watermark detectable without API access; negligible quality loss	Kirchenbauer *et al.*, arXiv 2023
Paraphrase detection baselines	Word-embedding + ML classifiers reach 0.72 F1 on academic corpora; string-matching services fare worse	Wahle & Gipp, arXiv 2023
Autoregressive Paraphrase Dataset	262 k aligned sentence pairs (original, paraphrase) with generator labels	Wahle *et al.*, HF Datasets

Table 1: Key prior work relevant to machine-paraphrased plagiarism.

2.1 Paraphrase Generation

Transformer paraphrasing began with fine-tuned T5 and BART models; today, zero-shot prompting of GPT-3.5/4 or Llama-2 dominates. Wahle et al.(2022) systematically evaluated LLM-facilitated plagiarism, showing that the detection precision of mainstream systems drops below 5%.

2.2 Watermarking LLM Output

Soft green-list (Kirchenbauer et al., 2024) — Randomly choose $G \subset V$ (size $\gamma|V|$); at decoding time boost logits of G by δ. The resulting token sequence reveals its origin via a skew in the distribution of G tokens. Other schemes encode explicit bitstrings (He et al., 2023) or embed watermarks in the prompt rather than the output (Bohacek et al., 2024). We choose the green list because of its favorable detection power for short windows.

3 Dataset

We adopt the *Autoregressive Paraphrase Dataset* (APD) released by Wahle *et al.*(2022). The default train split contains 262k examples spanning Wikipedia, theses, and arXiv domains. Each record holds `(text, label, model)`, where `label` $\in \{0, 1\}$ marks original versus paraphrase. We use 240k examples for training, 10k for validation, and 12k for testing; class balance is 50/50.

4 Methodology

We applied the watermark technique to the model, which is fine-tuned with the dataset, and we gave a sample of original text sentences from(Student theses, arXiv, Wikipedia), to generate watermark sentences on this dataset. Then the model generated a watermarked dataset, then we gave these watermarked sampled data to the Roberta-based model and generated 1-2 paraphrases on this dataset. The model gave results, so we tested clarity, fluency, and coherence compared with the original sentences. Then we used plagiarism detection tools and calculated F-1 scores for the tools and did analysis on the results.

4.1 Paraphrase Generation

A **T5-small** model (60 M parameters) is fine-tuned with prefix prompt "`paraphrase:`" for three epochs (batch 32, learning rate 3×10^{-5}, beam size 4).

4.2 Paraphrase Quality Scoring

To test the quality of the model-generated paraphrases compared to the original sentences, we used these parameters functions to compute:

- **Clarity** – sentence BLEU against original,
- **Fluency** – TextBlob sentiment polarity heuristic,
- **Coherence** – cosine similarity of [CLS] embeddings for RoBERTa-base.

4.3 Paraphrase Detection

We frame detection as binary classification using a **RoBERTa-base** encoder fine-tuned with cross-entropy loss. Baselines include Turnitin, Copyscape, and the zero-bit watermark detector of Kirchenbauer *et al.* Evaluation metric: F-1 score on the held-out set. Here is how to calculate F-1 score

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{1}$$

Where:
TP $=$ number of true positives
FP $=$ number of false positives
FN $=$ number of false negatives

5 Experiments and Results

We conducted some experiments and did some analysis with the data, and compared the data to get some insights. Here are our experiments:

1. We took a decoder model(**facebook/opt-1.3b**), tokenization done with the Auto tokenizer, and integrated with the watermark logits processor, with training arguments ($\gamma = 0.25, \delta = 2.0, self - hashseeding$), done fine tuning with the 262k dataset sampled, to this we applied the watermark technique so that whenever this model is given a sample dataset it will generate the watermarked sentences on the dataset.
2. Then we took a basic encoder model, RoBERTa and and we finetuned the model same and to this model we have done analysis with watermarked samples and without watermarked samples to test how the model is working. We took a sample of 1000 sentences (theses, Wikipedia, arXiv) and asked model to generate 1-2 paraphrases for each sentences, and we calculated parameters(clarity, coherence, fluency) to the paraphrased text, then we applied plagiarism detection tools and calculated F-1 scores for all these tools, then we gave this sample and generated watermarked sentences and we did the same process using this dataset. So here are our results for these experiments:

5.1 Paraphrase Quality

Metric	**Mean**	**SD**
Clarity (BLEU)	0.28	0.12
Fluency	0.34	0.08
Coherence	0.64	0.07

Table 2: Quality metrics on 1 000 generated sentences.

5.2 Detection Performance

This is the performance detection on the watermarked sample dataset. When we use these plagiarism detection tools, the results are as follows:

1. Here is the table for F-1 scores for the normal dataset for which the watermark technique is not applied, results:

Detector	F1
RoBERTa-base (itself)	**0.9479**
Turnitin	0.5058
Copyscape	0.5058

Table 3: F-1 scores held out on the normal dataset test sample.

2. Here is the table for F-1 scores for the watermarked dataset for which the watermark technique is applied, results:

Detector	F1
RoBERTa-base (itself)	**0.989**
Zero-bit Watermark	0.801
Turnitin	0.136
Copyscape	0.145

Table 4: F-1 scores held out on the watermarked dataset test sample.

5.3 Error Analysis

False positives chiefly occur in domain-specific jargon and equation-heavy sentences with low lexical overlap but genuine originality. False negatives cluster around short sentences ($\leq$15 tokens) lacking distinctive features.

6 Discussions

Three themes emerge:

1. **Transformer capacity matters** – even a compact RoBERTa succeeds when exposed to diverse generator styles.
2. **Commercial services lag** – pattern matching fails against paraphrases that retain semantics yet change surface form.
3. **Watermarking helps, but is not sufficient** – supervised training still outperforms unsupervised watermark detection.

Limitations include heuristic quality metrics, a lack of human evaluation, and a focus on English.

7 Conclusion and Future Work

We demonstrate that machine-paraphrased plagiarism is easily produced yet can be detected with near-perfect accuracy using a supervised Transformer classifier. Due to computational compatibility, we work on small models, but if we work on large models the results will be better enough to do evaluations and analysis. Future work will explore some uncertainty affecting outputs, and some uncertainty quantification methods to improve the model. Also, concentrating on multi-bit watermarks, multilingual corpora, and interpretability tools to assist academic integrity officers.

References

[1] J. P. Wahle, T. Ruas, F. Kirstein, and B. Gipp. *How Large Language Models are Transforming Machine-Paraphrased Plagiarism.* In *Proc. of EMNLP*, 2022.

[2] J. Kirchenbauer, J. Geiping, M. Goldblum, *et al.* *A Watermark for Large Language Models.* arXiv:2301.10226, 2023.

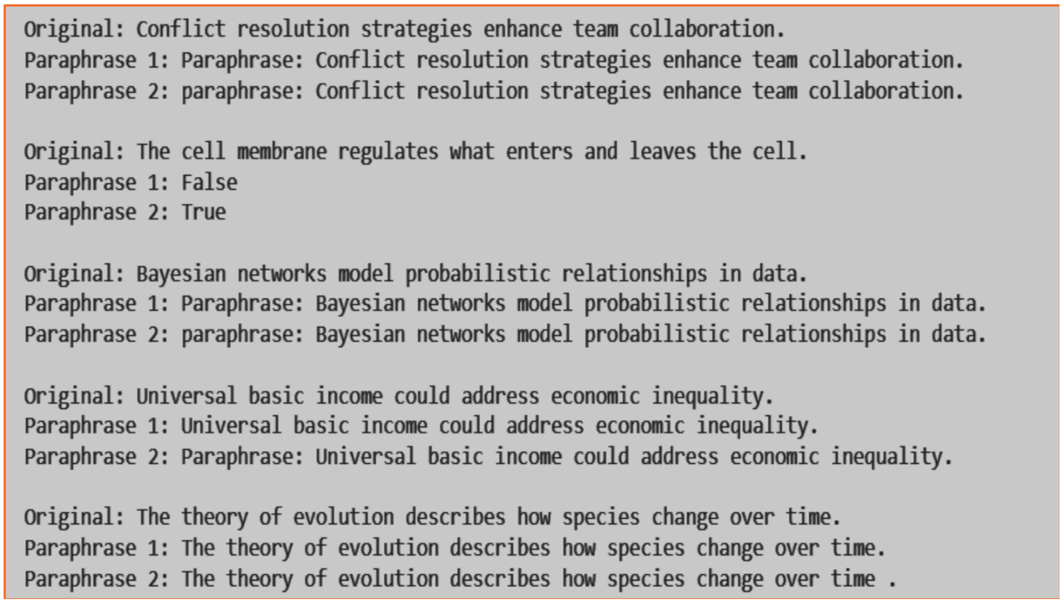
[3] J. P. Wahle and B. Gipp. *Identifying Machine-Paraphrased Plagiarism.* arXiv:2103.11909, 2023.

[4] J. P. Wahle, *et al.* Autoregressive Paraphrase Dataset. HuggingFace Datasets, 2023. `https://huggingface.co/datasets/jpwahle/autoregressive-paraphrase-dataset`

[5] J. De Guzman Fajardo. *Transformer Plagiarism Detection*. CS224n Final Report, Stanford University, 2024.

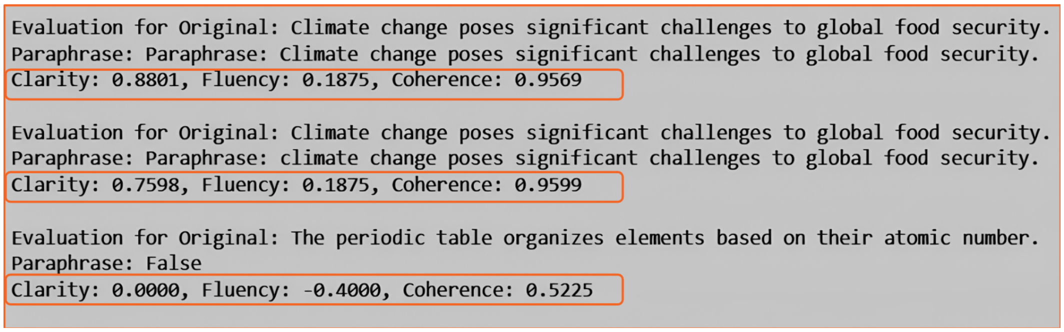
A Appendix

Here are the detailed explanations of our results. We have done some experiments mentioned already, so here is the detailed snapshots of our results, including paraphrase generation, evaluating the clarity, fluency, and coherence of the generated text for each sentences, and evaluating between the normal dataset and the watermarked dataset sample sentences.



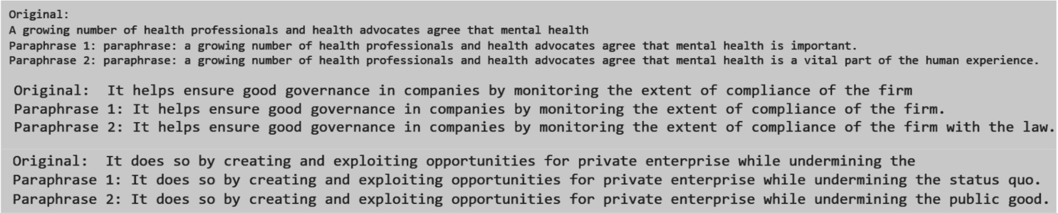
```
Original: Conflict resolution strategies enhance team collaboration.
Paraphrase 1: Paraphrase: Conflict resolution strategies enhance team collaboration.
Paraphrase 2: paraphrase: Conflict resolution strategies enhance team collaboration.

Original: The cell membrane regulates what enters and leaves the cell.
Paraphrase 1: False
Paraphrase 2: True

Original: Bayesian networks model probabilistic relationships in data.
Paraphrase 1: Paraphrase: Bayesian networks model probabilistic relationships in data.
Paraphrase 2: paraphrase: Bayesian networks model probabilistic relationships in data.

Original: Universal basic income could address economic inequality.
Paraphrase 1: Universal basic income could address economic inequality.
Paraphrase 2: Paraphrase: Universal basic income could address economic inequality.

Original: The theory of evolution describes how species change over time.
Paraphrase 1: The theory of evolution describes how species change over time.
Paraphrase 2: The theory of evolution describes how species change over time .
```

Figure 1: Model-generated paraphrases



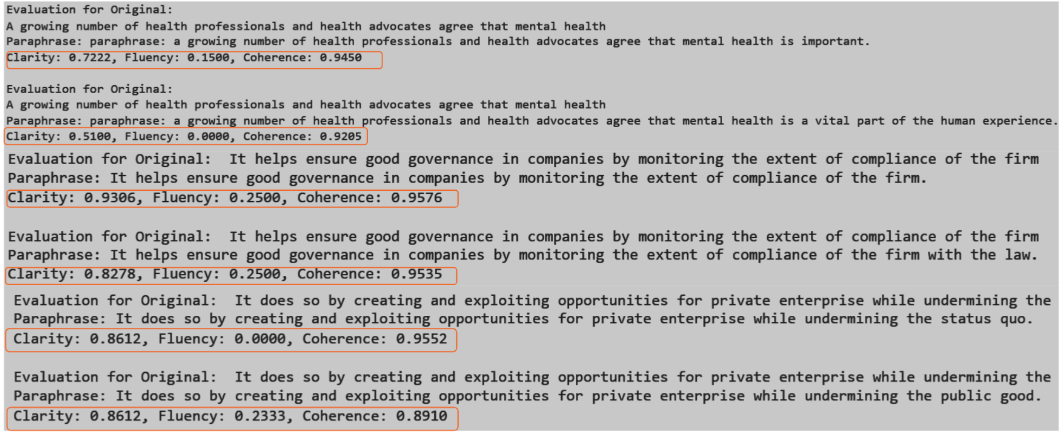
```
Evaluation for Original: Climate change poses significant challenges to global food security.
Paraphrase: Paraphrase: Climate change poses significant challenges to global food security.
Clarity: 0.8801, Fluency: 0.1875, Coherence: 0.9569

Evaluation for Original: Climate change poses significant challenges to global food security.
Paraphrase: Paraphrase: climate change poses significant challenges to global food security.
Clarity: 0.7598, Fluency: 0.1875, Coherence: 0.9599

Evaluation for Original: The periodic table organizes elements based on their atomic number.
Paraphrase: False
Clarity: 0.0000, Fluency: -0.4000, Coherence: 0.5225
```

Figure 2: Calculated clarity, fluency, and coherence for each model-generated sentence



```
Original:
A growing number of health professionals and health advocates agree that mental health
Paraphrase 1: paraphrase: a growing number of health professionals and health advocates agree that mental health is important.
Paraphrase 2: paraphrase: a growing number of health professionals and health advocates agree that mental health is a vital part of the human experience.

Original:  It helps ensure good governance in companies by monitoring the extent of compliance of the firm
Paraphrase 1: It helps ensure good governance in companies by monitoring the extent of compliance of the firm.
Paraphrase 2: It helps ensure good governance in companies by monitoring the extent of compliance of the firm with the law.

Original:  It does so by creating and exploiting opportunities for private enterprise while undermining the
Paraphrase 1: It does so by creating and exploiting opportunities for private enterprise while undermining the status quo.
Paraphrase 2: It does so by creating and exploiting opportunities for private enterprise while undermining the public good.
```

Figure 3: Model-generated paraphrases for the watermarked sample dataset sentences



```
Evaluation for Original:
A growing number of health professionals and health advocates agree that mental health
Paraphrase: paraphrase: a growing number of health professionals and health advocates agree that mental health is important.
Clarity: 0.7222, Fluency: 0.1500, Coherence: 0.9450

Evaluation for Original:
A growing number of health professionals and health advocates agree that mental health
Paraphrase: paraphrase: a growing number of health professionals and health advocates agree that mental health is a vital part of the human experience.
Clarity: 0.5100, Fluency: 0.0000, Coherence: 0.9205
Evaluation for Original:  It helps ensure good governance in companies by monitoring the extent of compliance of the firm
Paraphrase: It helps ensure good governance in companies by monitoring the extent of compliance of the firm.
Clarity: 0.9306, Fluency: 0.2500, Coherence: 0.9576

Evaluation for Original:  It helps ensure good governance in companies by monitoring the extent of compliance of the firm
Paraphrase: It helps ensure good governance in companies by monitoring the extent of compliance of the firm with the law.
Clarity: 0.8278, Fluency: 0.2500, Coherence: 0.9535
Evaluation for Original:  It does so by creating and exploiting opportunities for private enterprise while undermining the
Paraphrase: It does so by creating and exploiting opportunities for private enterprise while undermining the status quo.
Clarity: 0.8612, Fluency: 0.0000, Coherence: 0.9552

Evaluation for Original:  It does so by creating and exploiting opportunities for private enterprise while undermining the
Paraphrase: It does so by creating and exploiting opportunities for private enterprise while undermining the public good.
Clarity: 0.8612, Fluency: 0.2333, Coherence: 0.8910
```

Figure 4: Calculated clarity, coherence, and fluency for the watermarked sample