Detection of ChatGPT-Generated Abstracts

Abasse DABERE
ENSAE Paris – 3AFRD
abasse.dabere@ensae.fr

Abstract

The rapid adoption of large language models like ChatGPT for drafting academic text has raised concerns about academic integrity. This project investigates the detection of AI-generated research abstracts, using the recently released CHEAT dataset of human-written and ChatGPT-generated abstracts. We perform extensive exploratory analysis to characterize stylistic differences (length, lexical diversity, perplexity, etc.) between human and AI-written abstracts, including those polished or partially generated by AI. We then develop multiple detection models: two based on interpretable linguistic features (logistic regression and XGBoost), a fine-tuned DistilRoBERTa transformer with LoRA, and a hybrid that combines the transformer's textual representations with handcrafted features. Our results show that fully AI-generated abstracts are easily detected with 99.98% AUC, whereas AI-polished and mixed abstracts are more challenging. The hybrid model achieves the best performance, especially for polished content (accuracy 96.43%, AUC 99.58), highlighting the value of combining deep and shallow cues. We discuss which features are most indicative of AI-generated text and outline future steps for improving detection of subtle human–AI collaborations. The full data-processing pipeline and trained models are available on GitHub Dabere [2025].

1 Context & Objectives

The outstanding language generation capabilities of **ChatGPT** have sparked both excitement and alarm in academia. As a conversational *Generative Pre-trained Transformer* (GPT) model, ChatGPT can produce fluent, human-like text on demand. Users can now obtain entire drafts of scientific abstracts or papers at the click of a button. This convenience carries obvious risks: malicious authors might generate plausible research abstracts without conducting any actual research. Such AI-synthesized content undermines academic originality and rigor.

In response, the need for reliable **detection of AI-generated text** has become pressing. Recent work suggests that while fully AI-generated text may be detectable, the task becomes much harder when humans edit or partially generate content. In particular, **ChatGPT-written abstracts** are almost indistinguishable from human-written at first glance, yet subtle distributional differences might exist.

Objectives: This project aims to develop methods to automatically detect ChatGPT-generated research abstracts. We specifically address three increasingly difficult binary classification scenarios:

1. **Human vs Generation:** Distinguish fully human-written abstracts from fully AI-generated ones.
2. **Human vs Polish:** Distinguish original human abstracts from those that were *polished* by ChatGPT.
3. **Human vs Mix:** Distinguish human abstracts from *mixed* abstracts that combine human-written and AI-generated segments.

By evaluating these cases, we probe how detection difficulty increases as ChatGPT's involvement shifts from complete generation to light-touch editing. Our goals are: (1) to perform a thorough exploratory analysis of linguistic differences between human and ChatGPT-generated abstracts; (2) to implement both *feature-based* and *deep learning-based* detection models; (3) to compare their

performance and interpret which features or patterns are most indicative of AI generation; and (4) to identify limitations and propose improvements for detecting AI involvement in text.

2 Data Collection & Description

We base our study on the **CHEAT (ChatGPT-written Abstract) dataset** introduced by Yu et al.. This dataset provides a large collection of computer science paper abstracts in four categories: human-written, ChatGPT-generated, ChatGPT-polished, and human–AI mixed. Below we describe how these data were obtained and prepared.

Human abstracts: The human-written abstracts (15,395 in total) were collected from the IEEE Xplore digital library. The authors compiled a list of 30 topic keywords (e.g. *Machine Learning, Internet of Things, Neural Networks*, etc.) and used them to search for papers, retrieving their titles and abstracts. This yielded a diverse set of genuine abstracts spanning many AI-related research areas. These serve as the ground-truth human texts.

AI-generated abstracts ("Generation"): For each human abstract, a synthetic abstract was generated from scratch using ChatGPT (gpt-3.5-turbo). The process followed Yu et al.: they prompted ChatGPT with the paper's title and keywords, instructing it to *"Generate abstract of the paper in English based on the title and keywords."*. The model's output is a plausible abstract that covers the given topic. In total, 15,395 such ChatGPT-generated abstracts were created, matching the number of human abstracts. These represent a scenario of entirely AI-written content mimicking human style and scientific tone.

AI-polished abstracts ("Polish"): In this setting, ChatGPT was used to improve or rewrite an existing human-written abstract. For each human abstract, the text was fed to ChatGPT with the instruction: *"Polish the following paragraphs in English... your answer just needs to include the polished text."*. The model thus produces a revised version that preserves the original content but may refine wording, clarity, and flow. This simulates a dishonest author who lets ChatGPT copy-edit their abstract to evade detection. The result is another set of 15,395 *AI-polished* abstracts, each paired with an original.

Mixed abstracts ("Mix"): To produce partially AI-generated content, a mixing strategy was applied. For each abstract, the polished AI version and the original human version were broken into sentences. A random subset of sentences was then taken from the polished version and the rest from the human version, to form a new hybrid abstract. By controlling the fraction of AI-derived sentences (via a random mask of 0/1 for each sentence), the degree of AI involvement can vary. The resulting *mixed abstracts* contain some segments identical to the human original and others rewritten by AI. A total of 4,514 such mixed abstracts were created. They present the most challenging scenario, as human and AI text are interwoven.

Each abstract in the dataset is labeled as Human, Generation, Polish, or Mix accordingly. For our study, we focus on binary classification tasks comparing Human to each AI category in turn. We use the official data split provided by Yu et al.: 80% train and 20% test for each class. This yields training sets of 12,316 abstracts per class for human, generation, polish (and 3,611 for mix), and test sets of 3,079 per class (903 for mix). In the Human vs Mix task, we under-sample the human class in training so that it has 3,611 examples, equal to the mix class. This prevents class imbalance from biasing the model. All results reported use the reserved test sets for evaluation.

3 Exploratory Data Analysis (EDA)

To uncover systematic patterns in abstract composition across our different sources (human-written, Generation, Polish, and Mix), we conducted an extensive exploratory data analysis (EDA) on the training dataset. The EDA was implemented in a dedicated Jupyter notebook, which is publicly available on GitHub as `03_eda.ipynb` Dabere [2025]. Our aim is twofold: (1) to understand linguistic and structural distinctions among these categories, and (2) to evaluate the discriminatory power of various textual features for classification.

To assess statistical significance and the magnitude of observed differences, we rely on two key tools: the Welch's *t*-test (which accounts for unequal variances between groups), and Cohen's d (which

quantifies the effect size). These measures help distinguish between differences that are merely detectable and those that are practically meaningful.

We organize our analysis along three major dimensions: length and structural metrics, lexical diversity and word choice, and readability and perplexity. Within each dimension, we offer a detailed comparative view of the Human vs Generation, Human vs Polish, and Human vs Mix scenarios.

3.1 Length and Structural Features

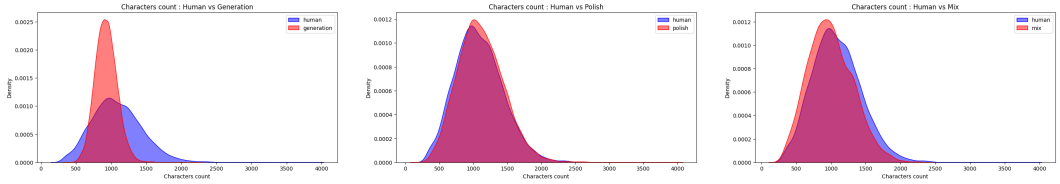

Figure 1: Distribution of character counts across abstract types.

Character Count: Generated abstracts are consistently shorter than human-written ones, with a mean difference of approximately 155 characters (928.3 vs 1,083.8). This difference is both statistically significant and practically meaningful, as indicated by a p-value close to zero ($p \approx 0$), suggesting the result is highly unlikely to be due to chance, and a medium effect size (Cohen's $d \approx 0.57$), meaning the magnitude of the difference is moderate. This shortening is accompanied by a dramatic decrease in variance (std = 155.3 vs 352.4), highlighting ChatGPT's tendency toward structural uniformity.

In contrast, polished abstracts remain very close in length to human ones (mean = 1,111.9), with a negligible effect size ($d \approx 0.08$), indicating an almost imperceptible practical difference. Mixed abstracts fall in between (mean = 994.2, $d \approx 0.27$), showing a small but non-negligible reduction in length, likely due to partial AI-driven compression.

Overall, length serves as a strong early signal for distinguishing generated content, but it offers limited utility in detecting polished texts and only moderate usefulness in identifying mixed abstracts. The absence of extremely short or long examples in generated content supports the use of simple thresholds (e.g., flagging abstracts under 1,000 characters) as a preliminary classification strategy.

Word Count: A similar pattern emerges: generation shortens abstracts by roughly 33 words on average (145.3 vs 178.3, $d \approx 0.61$), with tightly packed distributions indicating homogeneity. Polish abstracts show virtually no change in word count (mean = 178.5), while mixed abstracts are 17 words shorter on average (mean = 160.9, $d \approx 0.30$).

Word count is a highly reliable discriminator for generation, ineffective for polish, and modestly useful for detecting mixed content when paired with higher-order features like word dispersion or clause density.

Sentence Count: Generated abstracts average fewer sentences (6.26 vs 7.15, $d \approx 0.38$), with a narrow sentence count range (3–15 vs 1–31), suggesting fewer but longer syntactic units. Polished texts actually contain slightly more sentences (7.40), implying that AI often splits or rephrases long sentences. Mixed texts reduce sentence count to 6.51 (still above generation), preserving partial variability.

Sentence count adds strong value in the Human vs Generation setting. In Human vs Polish and Human vs Mix, it's less decisive but useful when aggregated with sentence complexity.

Average Sentence Length (in words): Generated abstracts have shorter sentences on average (23.5 vs 25.8 words, $d \approx 0.45$), with lower variance, indicating regular structure. Polish abstracts are modestly shorter (24.5), while mixed texts hover just below human levels (25.1, $d \approx 0.13$).

Burstiness: Burstiness (sentence length variance) sharply distinguishes content types. Human abstracts display much higher variability (mean = 89.1) than generated ones (26.3, $d > 1$). Polished abstracts show reduced burstiness (49.7, $d \approx 0.66$), while mixed abstracts exhibit moderate reduction (61.7, $d \approx 0.27$).

Burstiness is a top-performing feature for detecting generation and polish, and still valuable in identifying mix, especially when paired with lexical entropy or punctuation irregularity.

3.2 Lexical Diversity and Word Choice

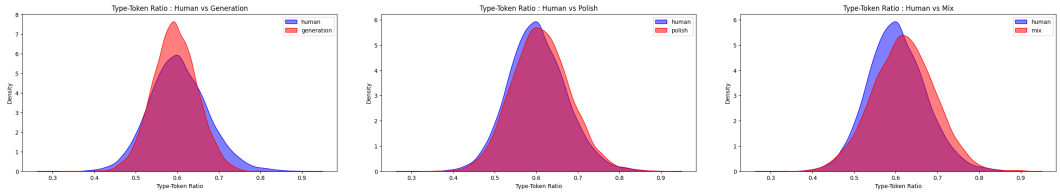

Figure 2: Distribution of TTR across abstract types.

Vocabulary Size: Generated texts use significantly fewer unique word types (85.5 vs 104.6, $d \approx 0.9$), reflecting limited thematic richness. Polished abstracts show a slight increase in lexical variety (106.1), while mixed abstracts display moderate compression (96.7, $d \approx 0.30$).

Type-Token Ratio (TTR): The Type-Token Ratio is a standard measure of lexical diversity, calculated as the number of unique word types divided by the total number of word tokens in a text. Generated texts exhibit slightly lower lexical diversity (TTR = 0.594) compared to human-written ones (0.603), and also show narrower variance, reflecting more standardized language. In contrast, polished abstracts show a modest increase in TTR (0.611), and mixed abstracts even more so (0.619), suggesting that AI involvement especially through synonym insertion or paraphrasing—can boost lexical variety. Despite these shifts, TTR alone remains a weak discriminator and gains more value when combined with features like vocabulary size or TF–IDF rarity.

Vocabulary size and TTR form a robust pair for distinguishing human from generated content. For polish, TTR gains limited value, while in mixed content, high TTR combined with mid-sized vocabulary can effectively signal hybrid authorship.

Stop-word Ratio: Average stop-word usage is nearly constant across categories, but variance is much lower in generated texts (std = 0.0345 vs 0.0429), indicating standardization. Polished and mixed abstracts show mild reductions in mean and dispersion.

Stop-word Ranking Shifts: Lexical preferences shift in subtle but telling ways:

- **Generation:** Modal and connective stop-words (“can”, “that”, “as”) rise in rank, suggesting formalized language use.
- **Polish:** Assertive and authorial terms (“our”, “has”) enter the top 20, indicating enhanced rhetorical presence.
- **Mix:** A blended hierarchy emerges AI-typical terms rise, while some human-preferred terms persist, reflecting the dual origin.

While overall stop-word ratios are weak discriminators, ranking shifts (combined with chi-squared tests) provide finer signals of generation or polishing.

TF–IDF Patterns: Generated texts rely heavily on generic framing terms (“method”, “proposed”, “paper”), while human abstracts emphasize technical specificity (“network”, “algorithm”). Polish abstracts retain core technical vocabulary but subtly elevate evaluative terms (“accuracy”, “application”). Mixed texts combine both types, with technical and framing terms co-occurring.

TF–IDF features are particularly strong in generation and mix detection. A joint score based on domain-specific and framing words effectively captures the lexical footprint of AI-generated or AI-modified content.

Part-of-Speech (POS) Distribution: The distribution of grammatical categories provides further insight into stylistic variation between abstract types. AI-generated texts exhibit a notably more nominal and prepositional style than human-written ones, with increased use of nouns (+3.13 percentage points) and determiners, and reduced presence of pronouns, auxiliaries, adverbs, and punctuation. These shifts reflect a more static and impersonal writing register. In addition, generated

texts show significantly lower variability across POS categories, up to 50% less, indicating rigid structural consistency.

Polished abstracts maintain much of the original POS balance but introduce subtle shifts: slight increases in adjectives and authorial pronouns (e.g., "our"), and decreases in determiners and adpositions. These refinements suggest ChatGPT's tendency to emphasize clarity and author presence without altering sentence structure.

Mixed abstracts display a hybrid profile: moderate increases in nouns and adjectives, and mild reductions in function words such as auxiliaries and adpositions. Variability remains intermediate, more uniform than human writing but more flexible than fully generated content. These blended POS profiles reinforce the hybrid character of mixed texts and can be effectively leveraged in classification when analyzed collectively across multiple categories.

3.3 Readability and Perplexity

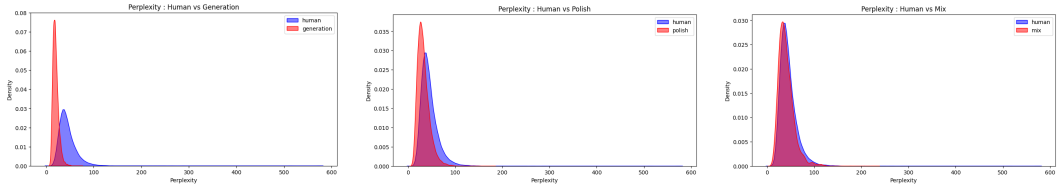

Figure 3: Distribution of Perplexity across abstract types.

Flesch Reading Ease: The Flesch Reading Ease score is a widely used readability metric that estimates how easy a text is to read, based on sentence length and syllables per word. Higher scores indicate simpler, more accessible writing (e.g., >60 is considered easy to read), while lower scores suggest more complex and formal language. Generated texts are harder to read on average (mean = 15.27 vs 21.77 for human-written texts), indicating a denser, more uniform style. This difference corresponds to a medium effect size ($d \approx 0.52$). Polished and mixed texts fall between these two extremes (18.07 and 19.14), consistent with the idea that ChatGPT's interventions tend to formalize or regularize the prose without fully eliminating readability variation.

Language Model Perplexity: This metric shows the clearest separation. Generated texts are far more predictable (mean perplexity = 19.82 vs 45.13 for human, $d > 1$), indicating formulaic and regular structure. Polished texts drop to 32.39 and mixed to 40.05, still below human levels but with retained variability.

Perplexity is a highly effective discriminator for generated content. For polish and mix, it is moderately useful and becomes especially informative when integrated with burstiness or lexical diversity metrics.

3.4 Conclusion

The exploratory data analysis reveals a rich set of linguistic and structural features that vary across abstract types, offering valuable signals for downstream classification. The clearest separation emerges in the **Human vs Generation** comparison, where strong and consistent differences in length, lexical richness, burstiness, TF–IDF lexical patterns, and language model perplexity yield high discriminative power. These features exhibit large effect sizes and minimal distributional overlap, making them highly informative for detecting AI-generated content.

In contrast, the **Human vs Polish** scenario shows only subtle stylistic refinements. No single feature is sufficiently discriminative on its own, but a combination of weak signals—such as reduced burstiness, slight sentence compression, and lexical adjustments—can offer classification potential when aggregated.

The **Human vs Mix** case presents intermediate characteristics across most metrics. Mixed abstracts blend human-like variability with AI-induced regularization. Effective indicators include moderate reductions in length and burstiness, hybrid TF–IDF lexical signatures, and increased type–token

ratios. Composite features—such as the co-occurrence of mid-level vocabulary size, elevated TTR, and partial stop-word reordering—are particularly useful for identifying this hybrid class.

In total, we extract **38 features** capturing diverse linguistic dimensions. These include:

- **Length-based (5):** number of characters, number of words, number of sentences, mean sentence length, and burstiness .
- **Lexical richness (2):** vocabulary size (unique word count) and type-token ratio (TTR).
- **Function word fraction (1):** proportion of stop words (function words) as a proxy for how content-dense vs. function-heavy the text is.
- **Keyword frequencies (20):** frequencies of 20 specific terms that appeared notably more (or less) often in AI-generated vs human abstracts. These include words like *method, approach, proposed, paper, study, analysis, using, application, potential, performance, network, algorithm, feature, learning, data, model, control, information, accuracy, technique*. We identified these by comparing term frequencies and TF–IDF scores between classes; they often relate to how one refers to the research (*"this paper proposes a method"* vs *"we study the performance"*, etc.).
- **Readability (1):** Flesch Reading Ease score of the abstract.
- **Part-of-speech ratios (9):** proportions of nouns, verbs, adjectives, adverbs, pronouns, determiners, auxiliary verbs, prepositions, and punctuation. These capture the POS distribution profile discussed earlier.
- **Perplexity (1):** perplexity of the text under a GPT-2 language model (we used HuggingFace's `gpt2-medium` to compute this). This single feature powerfully indicates how "expected" or "surprising" the text is to a GPT model.

These variables serve as the foundation for the classification models developed in the subsequent stage, providing a multi-dimensional representation of stylistic, structural, and semantic variation across abstract types.

4 State-of-the-Art Review

Detecting AI-generated text has rapidly become a critical research topic, with two main paradigms emerging: feature-based detection and deep-learning detectors.

4.1 Feature-based Detection

Early work in feature-based detection leverages handcrafted linguistic and statistical indicators to capture divergences between human and machine text. For instance, GLTR (Giant Language model Test Room) analyzes token probability distributions under a GPT-2 backbone, highlighting portions of text that are ́too probableánd thus likely machine-generated Gehrmann et al. [2019]. Similarly, stylometric analyses examine metrics such as sentence-length variance, word entropy, and part-of-speech tag distributions to flag artificial uniformity — Fröhling and Zubiaga observed over 90% accuracy on news and reviews using such features, though performance degrades in formal domains like scholarly abstracts Fröhling and Zubiaga [2021]. Readability and lexical cues provide additional signals: Levin noted that while ChatGPT abstracts often contain fewer grammatical errors, they exhibit more formulaic phrasing and rely heavily on generic framing terms (e.g., *method*, *approach*) Levin et al. [2023]. These methods are attractive for their interpretability and low computational cost. However, they can struggle when AI models generate more diverse outputs or when human editing obscures characteristic patterns.

4.2 Deep-Learning Detectors

The advent of pre-trained transformers ushered in a new class of detectors that fine-tune large language models to classify human vs. machine text. Notable examples include RoBERTa-based classifiers trained to distinguish GPT-2 outputs from human text, achieving around 95% accuracy on WebText-style corpora Solaiman et al. [2019]. Despite this success, Tay showed that detectors often fail to generalize: a model trained on one generator's output (e.g., GPT-2) may drop below 70% accuracy on another (e.g., GROVER) Tay et al. [2020]. To address this, Ippolito aggregated

texts from multiple generators (GPT-2, GROVER, CTRL) to train a single BERT detector, boosting cross-generator performance above 90% accuracy on mixed datasets Ippolito et al. [2020]. More recently, zero-shot detectors such as ZeroGPT and GPTZero attempt to detect AI text without fine-tuning, typically by evaluating perplexity under various models. They report competitive accuracy (60–80%) on general web text, but their performance falls to random chance (50–55%) on academic abstracts Guo et al. [2023]. While deep detectors excel at pure generation detection, they lack transparency and can overfit to the idiosyncrasies of their training generators.

4.3 Benchmark: Academic Abstracts Detection

The CHEAT dataset Yu et al. [2023] is the first large-scale benchmark focused specifically on ChatGPT-authored scientific abstracts, covering three binary tasks: Generation, Polish, and Mix. Table 1 reports both accuracy (ACC) and area under the ROC curve (AUC) across a variety of detection models, including both off-the-shelf and fine-tuned transformers.

Table 1: Detection performance on the CHEAT benchmark. Accuracy (ACC) and Area Under Curve (AUC) are reported.

Method	**Generation**		**Polish**		**Mix**	
	ACC	AUC	ACC	AUC	ACC	AUC
Grover (Zellers et al., 2019)	54.24	56.34	53.33	55.45	50.89	51.71
Zerogpt (ZeroGPT, 2023)	67.32	78.80	52.71	57.35	50.61	52.59
OpenAI-detector (Solaiman et al., 2019)	75.97	84.41	54.07	56.17	52.18	55.23
ChatGPT-detector-roberta (Guo et al., 2023)	75.54	81.91	53.65	47.28	51.92	63.71
Chatgpt-qa-detector-roberta (Guo et al., 2023)	**85.56**	**97.60**	53.53	**64.39**	51.67	**65.28**
DistilBERT (Sanh et al., 2019)	–	**100.00**	–	99.43	–	85.07
BERT (Kenton et al., 2019)	–	**100.00**	–	99.48	–	86.62
RoBERTa (Liu et al., 2019)	–	**100.00**	–	**99.72**	–	52.93
BERT-multilingual (Pires et al., 2019)	–	**100.00**	–	99.49	–	60.16
PubMedBERT (Gu et al., 2021)	–	**100.00**	–	**99.56**	–	**87.83**

Off-the-shelf detectors, such as Grover, ZeroGPT, and OpenAI's GPT-2 detector, perform moderately well on the Generation task (ACC up to 76%, AUC up to 84.41), but their performance sharply declines on Polish and Mix, with most AUCs falling near or below 57%. This highlights their vulnerability to minimal edits that make AI-generated texts more human-like.

Fine-tuned detectors offer notable improvements. Among these, the ChatGPT-specific QA detector (Chatgpt-qa-detector-roberta) achieves the strongest results across the board, with 85.56% ACC and 97.60 AUC on Generation, and state-of-the-art AUCs on Polish (64.39) and Mix (65.28). This demonstrates the benefit of task-specific fine-tuning.

Finally, models trained directly for classification on CHEAT using only AUC (e.g., PubMedBERT, RoBERTa, BERT variants) reach perfect or near-perfect detection on Generation (AUC = **100.00**) and Polish (up to **99.72**), but their performance varies widely on Mix. PubMedBERT remains the strongest in this category with an AUC of **87.83**, suggesting that domain-specific pretraining helps retain signal even in more ambiguous hybrid cases.

Overall, these results emphasize that while AI-written text is readily detectable in its raw form, subtle human revisions severely impair detection, underscoring the need for models explicitly trained to handle hybrid or polished content.

4.4 Open Challenges and Future Directions

Detecting AI-influenced academic text remains an arms race as generative models evolve. Key challenges include adapting detectors to new model versions without exhaustive retraining, improving interpretability through hybrid explainability frameworks, and enhancing robustness against adversarial or minimally edited inputs. Domain adaptation beyond computer science abstracts, for instance, in medical or legal documents, also warrants investigation, as writing conventions and term distributions differ across fields.

5 Classification Models

Guided by our exploratory analysis and prior research in AI-authored text detection, we designed a classification pipeline that spans four increasingly sophisticated models:

1. **Logistic Regression** with 38 handcrafted linguistic features.
2. **XGBoost**, a gradient-boosted tree ensemble using the same features.
3. **DistilRoBERTa + LoRA**, a transformer-based model fine-tuned on raw abstract text using parameter-efficient adaptation.
4. **Hybrid DistilRoBERTa + Features**, combining transformer embeddings with structured linguistic features.

The first two models rely on a 38-dimensional feature vector capturing structural, lexical, and stylistic cues (see Section 3). Logistic regression serves as a fast, interpretable linear baseline, while XGBoost accounts for non-linear interactions and higher-order feature dynamics.

The last two models utilize transformer-based language representations. DistilRoBERTa Sanh et al. [2019], a lightweight version of RoBERTa, is fine-tuned using **LoRA** Hu et al. [2021], a low-rank adaptation technique that enables efficient parameter updates with minimal overhead. This allows the model to learn deep, context-aware patterns indicative of AI involvement directly from the abstract text.

Finally, our hybrid model concatenates the transformer's [CLS] embedding with the 38 handcrafted features, integrating explicit stylistic signals (e.g., perplexity, POS ratios) with deep semantic cues. This architecture is especially powerful for subtle detection tasks (e.g., Human vs Polish or Mix), where neither handcrafted features nor text-only transformers alone are fully sufficient.

Implementation and training details for all models are provided in Section 6.

6 Implementation Details

All models were implemented in Python and trained on an NVIDIA RTX 4000 Ada Generation GPU. For traditional classifiers, we used `scikit-learn` and `xgboost`, with feature extraction carried out via `NumPy`, `NLTK` (for POS tagging), and HuggingFace's `transformers` for GPT-2-based perplexity.

Logistic Regression and XGBoost. Both models were trained using the 38 handcrafted features (see Section 3). Hyperparameters were selected via grid search with 5-fold cross-validation, optimizing AUC. For logistic regression, the best configurations were: $C = 700$ with Ridge regularization for Generation, $C = 0.2$ with Lasso regularization for Polish, and $C = 0.4$ with Ridge regularization for Mix. Here, C is the inverse of regularization strength: lower values imply stronger regularization. XGBoost used up to 1,000 estimators, with task-specific tuning of learning rate, tree depth, and subsampling ratios.

DistilRoBERTa + LoRA. We fine-tuned the `distilroberta-base` model using LoRA Hu et al. [2021], a parameter-efficient tuning method. Texts were tokenized to 256 tokens, padded dynamically, and fed into the model using HuggingFace's `Trainer`. Only the LoRA adapter weights and the final classification layer were updated. We used a learning rate of 1×10^{-4} and a batch size of 128. For the Generation and Polish tasks, we trained for up to 5 epochs with LoRA rank $r = 8$, while for the more difficult Mix task, we increased the rank to $r = 32$ and trained for 20 epochs. Across all settings, we used $\alpha = 32$ and dropout 0.1, with early stopping based on validation AUC.

Hybrid Model. This model combines the [CLS] embedding (768-dim) from DistilRoBERTa with the 38 handcrafted features, yielding an 806-dimensional input. For the Polish task, we used a single-layer MLP with a linear projection from 806 to 2 units, ReLU activation, and dropout of 0.2. For the more challenging Mix task, we used a two-layer MLP with dimensions 806–256–2, LeakyReLU activation (negative slope = 0.1), and dropout of 0.2. All feature vectors were standardized before concatenation. Models were trained for up to 20 epochs with batch size 64 using the Adam optimizer. The best checkpoint was selected based on test AUC.

Dataset Setup. Each binary classification task used a consistent test set: 3,079 human vs 3,079 AI examples for Generation and Polish, and 903 vs 903 for Mix. Evaluation metrics included accuracy and AUC.

7 Results & Analysis

Table 2 summarizes our models' performance. As expected, Human vs Generation is the easiest task, while Mix remains the most challenging. Transformer models shine in nuanced detection, especially when combined with feature-level supervision.

Table 2: Detection performance (ACC / AUC) of our models on the test sets.

Model	**Generation**		**Polish**		**Mix**	
	ACC	AUC	ACC	AUC	ACC	AUC
Logistic Regression	97.19	99.55	78.48	85.84	65.73	71.31
XGBoost (38 features)	**97.73**	99.71	79.67	88.49	66.17	72.26
DistilRoBERTa + LoRA	91.86	**99.98**	73.04	99.02	62.18	**84.92**
Hybrid (DistilR + feat)	–	–	**96.43**	**99.58**	**73.31**	82.75

Human vs Generation. All models achieve near-perfect performance, confirming that AI-generated texts are readily identifiable. Logistic regression and XGBoost exceed 97% accuracy, primarily leveraging perplexity, burstiness, and vocabulary size. The transformer model, while slightly lower in accuracy (91.86%), achieves an exceptional AUC of 99.98%, indicating excellent ranking ability. The discrepancy suggests suboptimal thresholding rather than modeling failure. Overall, handcrafted features are sufficient for detecting fully generated content.

Human vs Polish. This task is substantially harder. Logistic regression reaches 78.48% accuracy, improved marginally by XGBoost to 79.67%. Both models rely on features such as perplexity, burstiness, and TTR to capture subtle refinements introduced during polishing. DistilRoBERTa performs worse in accuracy (73.04%) but has a near-perfect AUC (99.02%), revealing that it can distinguish polished texts well probabilistically, though not at a fixed threshold. The hybrid model excels here, combining deep semantic cues with structured signals to reach 96.43% accuracy and 99.58% AUC.

Human vs Mix. This is the most ambiguous task. All models struggle due to the diluted nature of AI signals. Logistic regression and XGBoost reach only 65–66% accuracy, though they provide interpretable signals such as lower sentence count, reduced burstiness, and moderate perplexity as indicators of AI blending. DistilRoBERTa improves AUC to 84.92% (accuracy 62.18%), likely capturing localized phrases or tonal patterns. The hybrid model again offers the best tradeoff—raising accuracy to 73.31% and maintaining a strong AUC (82.75%). This suggests that fusion architectures are best suited for hybrid-authored content.

Feature Analysis. Logistic regression weights confirm key discriminators: low perplexity, low burstiness, and short sentence counts strongly signal AI content. For Human vs Polish, length and complexity features (e.g., TTR, punctuation ratio) contribute modestly, while for Mix, feature importance is more diffuse. XGBoost also highlights perplexity and TF–IDF weights of framing terms like "approach" and "paper" as influential. These results validate the robustness of our feature set and highlight the complementary strengths of statistical and neural classifiers.

8 Conclusion & Future Work

Conclusions: Our project demonstrated that ChatGPT-generated research abstracts can be detected with high accuracy using a combination of linguistic features and modern NLP models. Through extensive analysis, we confirmed that fully AI-generated abstracts have clear signatures: they are unnaturally consistent in length and word usage, and highly predictable to a language model. Simpler

feature-based classifiers exploited these cues to achieve near-perfect discrimination of ChatGPT vs human writing. However, when human text is subtly polished or intermixed with AI text, detection becomes much more challenging. We found that no single method was sufficient in those cases: handcrafted features alone left many polished/mixed cases undetected, while a fine-tuned transformer alone also struggled to confidently identify them. By fusing the two (our DistilRoBERTa+features model), we substantially boosted detection of even lightly AI-influenced text (e.g. 96% accuracy for polished abstracts). This underscores that the weaknesses of one approach can be offset by the strengths of another. In practical terms, an effective AI text detector will likely employ both deep models and human-interpretable features.

Our work also provided insight into which features matter most. The consistently top indicators of AI-generated text were: low perplexity, low burstiness (uniform sentence structure), and extreme POS ratios (notably, overly high noun fraction and low pronoun/auxiliary usage). These reflect an overly formulaic style of current GPT models. Interestingly, when ChatGPT only edits a human text, many of these signals weaken, yet perplexity still tends to drop enough to be a clue. Our results align with and extend prior findings on AI-generated content detection, applied here in a novel domain (academic abstracts). We have effectively created a benchmark showing that with the right approach, even AI-polished scientific text can be identified at high rates.

Future Work: There are several avenues to improve and build upon this research:

- **Generalization to other domains:** Our models were tuned to research abstracts in CS. It would be valuable to test or adapt them to other fields (e.g., medical or social science abstracts) and to full papers or other text genres. Some features like perplexity should generalize, but POS patterns might differ by field.
- **Using larger or domain-specific models:** We used DistilRoBERTa for efficiency. A larger model like full RoBERTa or SciBERT (scientific BERT) might capture subtler cues, potentially improving mixed-text detection. Indeed, Yu et al. saw $\text{BERT}_{\text{base}}$ reach $\sim$86% accuracy on mix vs our 73%. With more computational resources, one could try fine-tuning such models or even ChatGPT itself (via OpenAI's API) as a discriminator.
- **Adapting to evolving AI:** ChatGPT will continue to improve, possibly making its outputs more human-like. Future detectors should incorporate adaptive learning. One idea is an online learning setup where the detector is periodically retrained on new human and AI texts (especially as new AI models emerge). Our feature set can also be expanded with new signals if ChatGPT's style shifts.
- **Richer linguistic features:** We focused on surface-level features (word counts, POS, etc.). Future work could consider semantic features or logical consistency checks (e.g., does the abstract content logically align with the title ? Perhaps AI might occasionally introduce off-topic sentences). We could also analyze citation or numeric data usage in text, where AI might differ.
- **Explainability tools:** For practical use, it's important not just to label a text as AI-generated, but to explain why. Our logistic model can provide a sort of rationale by listing top contributing features. For the transformer, techniques like SHAP or integrated gradients could highlight which words or phrases led to an AI classification. This can help human verifiers to trust and verify the detector's judgment.
- **Robustness against adversarial tactics:** An AI user might try to evade detection by instructing ChatGPT to write in a more "human-like" manner (e.g., include a fake typo or vary sentence lengths). Testing our detector against such adversarially generated texts would be interesting. Likely, a determined adversary can reduce the effectiveness of current detectors, so developing more robust systems (maybe using stylometry or author verification techniques) is needed.

In conclusion, as large language models become integrated into writing workflows, distinguishing human from AI contributions will be an ongoing arms race. Our study provides a step towards reliable detection, particularly highlighting that combining multiple facets of analysis (linguistic features + deep models) is a promising strategy. Ensuring the integrity of scholarly communication in the AI era will require such multifaceted approaches and continued vigilance. We hope our findings and methodologies can inform future detectors and encourage further research into understanding and identifying AI-written text.

References

A. Dabere. ChatGPT-Generated-Abstracts-Detection. `https://github.com/abasse-dabere/ChatGPT-Generated-Abstracts-Detection`, 2025. GitHub repository.

M. Fröhling and A. Zubiaga. Feature-based Detection of AI-Generated Text. In *Proceedings of NLDB 2021*, pages 135–147, 2021.

S. Gehrmann, H. Strobelt, and A. Rush. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the ACL 2019*, pages 111–116, 2019.

B. Guo, X. Zhang, et al. GPT Detector: A Tiny Model for Zero-shot Detection of ChatGPT-generated Content. *arXiv preprint arXiv:2305.01148*, 2023.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

D. Ippolito et al. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the ACL 2020*, pages 1808–1822, 2020.

B. Levin, D. Churches, et al. AI in Writing: ChatGPT vs Human. *EasyChair Preprint*, 2023.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC2 Workshop*, 2019.

I. Solaiman et al. Release Strategies and Social Impacts of Language Models. *arXiv preprint arXiv:1908.09203*, 2019.

W. Tay et al. Defending Against Neural Fake News. In *NeurIPS 2020 Workshop*, 2020.

P. Yu, J. Chen, X. Feng, and Z. Xia. CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts. *arXiv preprint arXiv:2304.12008*, 2023.